PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-348061
(43)Date of publication of application : 15.12.2000

(51)Int.Cl. G06F 17/30
G06F 12/00

(21)Application number : 11-162990
(22)Date of filing : 09.06.1999

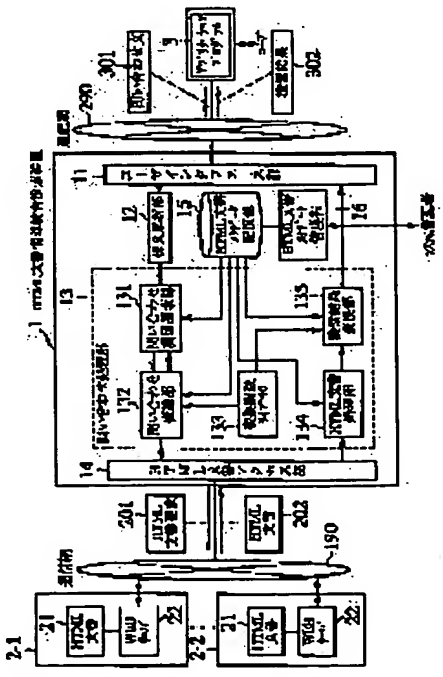
(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>
(72)Inventor : IIZUKA YUICHI
TSUNAKAWA MITSUAKI
NAGAMATSU HISAHIRO
HOSHINO TAKASHI
MACHIHARA HIROKI

(30)Priority

Priority number	Priority date	Priority country
10162648	10.06.1998	JP
10219365	03.08.1998	
11096183	02.04.1999	JP
		JP

(54) SEMI-STRUCTURED DOCUMENT INFORMATION INTEGRATING RETRIEVAL DEVICE, SEMI-STRUCTURED DOCUMENT INFORMATION EXTRACTING DEVICE, ITS METHOD AND RECORDING MEDIUM FOR STORING ITS PROGRAM

(57)Abstract:
PROBLEM TO BE SOLVED: To realize batch item unit retrieval concerning internal information over a plurality of semi-structured documents which are scattered on an open network.
SOLUTION: An integrating retrieval device is constituted by providing a document position locating part 131 for obtaining the position of the semi- structured documents based on an input inquiry, an inquiry converting part 132 for converting an inquiry into the expression form of an item corresponding to a retrieval item in the semi-structured document, a document retrieving part 14 for obtaining the semi-structure documents by the converted inquiry, a document processing part 134 for extracting item data based on document structure information for division at every item to be extracted, selecting the extracted item data based on attribute information for retrieving the item by condition and adopting it as a retrieval result and a retrieval result converting part 135 for converting the retrieval result into the expression form of the item which is defined at every user.



LEGAL STATUS

[Date of request for examination]	09.06.1999
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	3160265
[Date of registration]	16.02.2001
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－348061
（Ｐ2000－348061Ａ）
(43)公開日　平成12年12月15日（2000.12.15）

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'（参考）
Ｇ０６Ｆ　17/30		Ｃ０６Ｆ　15/40	３４０	５Ｂ０７５
12/00	５４６	12/00	５４６Ｒ	５Ｂ０８２
		15/40	３１０Ｆ	
		15/419	３２０	

審査請求　有　請求項の数37　ＯＬ　（全 51 頁）

(21)出願番号　特願平11－162990

(22)出願日　平成11年６月９日（1999.6.9）

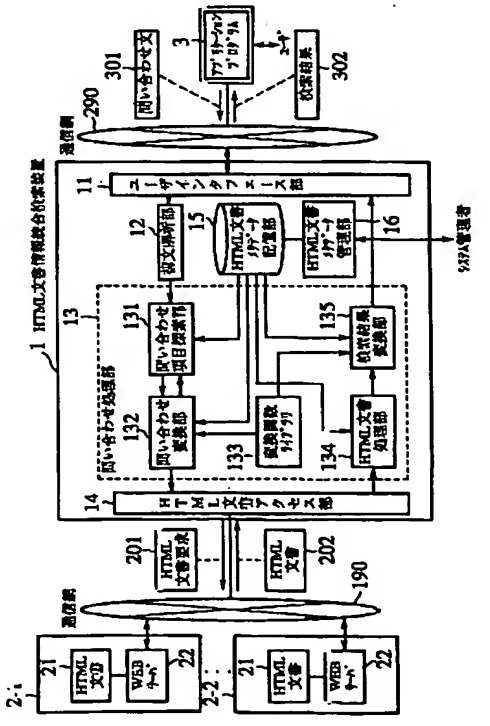
(31)優先権主張番号　特願平10－162648
(32)優先日　平成10年６月10日（1998.6.10）
(33)優先権主張国　日本（ＪＰ）
(31)優先権主張番号　特願平10－219365
(32)優先日　平成10年８月３日（1998.8.3）
(33)優先権主張国　日本（ＪＰ）
(31)優先権主張番号　特願平11－96183
(32)優先日　平成11年４月２日（1999.4.2）
(33)優先権主張国　日本（ＪＰ）

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(72)発明者　飯塚　裕一
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(72)発明者　綱川　光則
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(74)代理人　100083806
弁理士　三好　秀和　（外１名）

最終頁に続く

(54)【発明の名称】　半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびにそのプログラムを格納する記録媒体

(57)【要約】
【課題】　オープンなネットワーク上に散在する複数の半構造化文書に跨って内在する情報への一括した項目単位の検索を実現する。
【解決手段】　入力問い合わせに基づき半構造化文書の所在を得る文書所在探索部１３１と、問い合わせを半構造化文書中の検索項目に対応する項目の表現形式に変換する問い合わせ変換部１３２と、変換された問い合わせにより半構造化文書を取得する文書検索部１４と、半構造化文書から、抽出すべき項目ごとに区切るための文書構造情報に基づいて、項目データを抽出し、項目を条件検索するための属性情報に基づいて前記抽出された項目データを選択して検索結果とする文書処理部１３４と、検索結果を、各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部１３５とを具備して統合検索装置を構成する。



【特許請求の範囲】
【請求項１】　オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索装置であって、
半構造化文書ごとに該半構造化文書中から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報を記憶する記憶部と、
入力された問い合わせから、前記メタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得る検索部と、
ユーザーごとに所定の単一フォーマットで前記検索結果を出力する出力部とを具備することを特徴とする半構造化文書情報統合検索装置。
【請求項２】　オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索装置であって、
オープンネットワーク上での半構造化文書の所在を示す所在情報と、前記半構造化文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報と、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶する記憶部と、
検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を前記所在情報から得る文書所在探索部と、
入力された前記問い合わせを、必要に応じ、前記表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換する問い合わせ変換部と、
前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得する文書検索部と、
取得された各半構造化文書から、前記文書構造情報に基づいて、項目データを抽出し、必要に応じて前記検索条件を用い、前記属性情報に基づいて前記抽出された項目データを選択して検索結果とする文書処理部と、
前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部とを具備することを特徴とする半構造化文書情報統合検索装置。
【請求項３】　上記半構造化文書情報統合検索装置は、さらに、
半構造化文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、半構造化文書から抽出すべき項目群の所定の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部を具備し、
前記文書処理部は、前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応する前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とすることを特徴とする請求項２に記載の半構造化文書情報統合検索装置。
【請求項４】　前記文書処理部は、前記検索結果を、表形式に整形することを特徴とする請求項３に記載の半構造化文書情報統合検索装置。
【請求項５】　前記文書処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他の半構造化文書へのリンク情報を含む場合には、リンク先の半構造化文書をスキャンして、前記リンク先の半構造化文書と前記テンプレートとを比較することを特徴とする請求項３に記載の半構造化文書情報統合検索装置。
【請求項６】　前記テンプレートは、半構造化文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項３に記載の半構造化文書情報統合検索装置。
【請求項７】　前記テンプレートは、半構造化文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、
前記文書処理部は、前記取得された半構造化文書をスキャンして、該半構造化文書の前記部分構造と、該半構造化文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項３に記載の半構造化文書情報統合検索装置。
【請求項８】　前記テンプレートは、半構造化文書が互いに異なる要素からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項３に記載の半構造化文書情報統合検索装置。
【請求項９】　オープンネットワーク上の複数のサーチエンジンを介して情報を検索する半構造化文書情報統合検索装置であって、
オープンネットワーク上でのサーチエンジンの所在を示す所在情報と、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報と、ＨＴＭＬ文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報と、ユーザーの項目の表現形式と各ＨＴＭＬ文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶する記憶部と、
検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を前記所在情報から得る文書所在

探索部と、
前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件を満たす検索条件が指定されたサーチエンジンを、前記入力必須項目情報から得て、検索対象サーチエンジンとして選択するサーチエンジン選択部と、
前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定部と、
前記変換された問い合わせ群のそれぞれを、必要に応じ、前記表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換部と、
前記変換された問い合わせを前記得られた所在に送信して、ＨＴＭＬ文書を取得する文書検索部と、
各サーチエンジンから取得されたＨＴＭＬ文書からなる第１の検索結果から、前記文書構造情報に基づいて、項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用い、前記属性情報に基づいて、前記抽出された項目データを選択して、第２の検索結果とする検索結果処理部と、
前記第２の検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部とを具備することを特徴とする半構造化文書情報統合検索装置。
【請求項１０】　上記半構造化文書情報統合検索装置は、さらに、
ＨＴＭＬ文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部を具備し、
前記文書処理部は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、第２の検索結果とすることを特徴とする請求項９に記載の半構造化文書情報統合検索装置。
【請求項１１】　前記文書処理部は、前記検索結果を、表形式に整形することを特徴とする請求項１０に記載の半構造化文書情報統合検索装置。
【請求項１２】　前記文書処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他のＨＴＭＬ文書へのリンク情報を含む場合には、リンク先のＨＴＭＬ文書をスキャンして、前記リンク先のＨＴＭＬ文書と前記テンプレートとを比較することを特徴とする請求項１０に記載の半構造化文書情報統合検索装置。
【請求項１３】　前記テンプレートは、ＨＴＭＬ文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項１０に記載の半構造化文書情報統合検索装置。
【請求項１４】　前記テンプレートは、ＨＴＭＬ文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、
前記文書処理部は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項１０に記載の半構造化文書情報統合検索装置。
【請求項１５】　前記テンプレートは、ＨＴＭＬ文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項１０に記載の半構造化文書情報統合検索装置。
【請求項１６】　オープンネットワーク上の任意のＨＴＭＬ文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出装置であって、
ＨＴＭＬ文書ごとに、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部と、
取得されたＨＴＭＬ文書に対応するテンプレートを解析するテンプレート解析部と、
前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理部とを具備することを特徴とする半構造化文書情報抽出装置。
【請求項１７】　前記テンプレート処理部は、前記抽出された項目データを、表形式に整形することを特徴とする請求項１６に記載の半構造化文書情報抽出装置。
【請求項１８】　前記テンプレート処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他のＨＴＭＬ文書へのリンク情報を含む場合には、リンク先のＨＴＭＬ文書をスキャンして、前記リンク先のＨＴＭＬ文書と前記テンプレートとを比較することを特徴とする請求項１６に記載の半構造化文書情報抽出装置。
【請求項１９】　前記テンプレートは、ＨＴＭＬ文書が複数の同一部分構造を有する場合、各部分構造ごとに抽

出テキスト形式情報が記述され、
前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項１６に記載の半構造化文書情報抽出装置。
【請求項２０】　前記テンプレートは、ＨＴＭＬ文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、
前記テンプレート処理部は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項１６に記載の半構造化文書情報抽出装置。
【請求項２１】　前記テンプレートは、ＨＴＭＬ文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項１６に記載の半構造化文書情報抽出装置。
【請求項２２】　オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索方法であって、
入力された問い合わせから、半構造化文書ごとに該半構造化文書から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得るステップと、
ユーザーごとに所定の単一フォーマットで前記検索結果を出力するステップとを含むことを特徴とする半構造化文書情報統合検索方法。
【請求項２３】　オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索方法であって、
検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を、オープンネットワーク上での半構造化文書の所在を示す所在情報から得るステップと、
入力された前記問い合わせを、必要に応じ、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換するステップと、
前記変換された検索要求を前記得られた所在に送信して、半構造化文書を取得するステップと、
取得された各半構造化文書から、半構造化文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて前記検索条件を用い、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報に基づいて、前記抽出された項目データを選択して検索結果とするステップと、
前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換するステップとを含むことを特徴とする半構造化文書情報統合検索方法。
【請求項２４】　オープンネットワーク上の複数のサーチエンジンを介して情報を検索する半構造化文書情報統合検索方法であって、
検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を、オープンネットワーク上でのサーチエンジンの所在を示す所在情報から得るステップと、
前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件が指定されたサーチエンジンを、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報から得て、検索対象サーチエンジンとして選択するステップと、
前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換するステップと、
前記変換された問い合わせ群のそれぞれを、必要に応じ、ユーザーの項目の表現形式と各ＨＴＭＬ文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換するステップと、
前記変換された問い合わせを前記得られた所在に送信して、ＨＴＭＬ文書を取得するステップと、
各サーチエンジンから取得されたＨＴＭＬ文書からなる第１の検索結果から、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用いて項目を条件検索するための属性を規定する属性情報に基づいて前記抽出された項目データを選択して、第２の検索結果とするステップと、
前記第２の検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換するステップとを含むことを特徴とする半構造化情報統合検索方法。
【請求項２５】　オープンネットワーク上の任意のＨＴＭＬ文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出方法であって、

取得されたＨＴＭＬ文書に対応し、ＨＴＭＬ文書ごとに、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを解析するステップと、
前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するステップとを含むことを特徴とする半構造化文書情報抽出方法。
【請求項２６】　オープンネットワーク上の複数の半構造化文書に内在する情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、
入力された問い合わせから、半構造化文書ごとに該半構造化文書から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得る処理と、
ユーザーごとに所定の単一フォーマットで前記検索結果を出力する処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体。
【請求項２７】　オープンネットワーク上の複数の半構造化文書に内在する情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、
検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を、オープンネットワーク上での半構造化文書の所在を示す所在情報から得る文書所在探索処理と、
入力された前記問い合わせを、必要に応じ、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換する問い合わせ変換処理と、
前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得する文書検索処理と、
取得された各半構造化文書から、半構造化文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて前記検索条件を用い、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報に基づいて、前記抽出された項目データを選択して検索結果とする検索結果生成処理と、
前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体。
【請求項２８】　前記検索結果生成処理は、
前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応し、半構造化文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、半構造化文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とすることを特徴とする請求項２７に記載のコンピュータ読み取り可能な記録媒体。
【請求項２９】　前記検索結果生成処理は、前記検索結果を、表形式に整形することを特徴とする請求項２８に記載のコンピュータ読み取り可能な記録媒体。
【請求項３０】　前記検索結果生成処理は、前記テンプレート中の前記抽出テキスト形式情報が、他の半構造化文書へのリンク情報を含む場合には、リンク先の半構造化文書をスキャンして、前記リンク先の半構造化文書と前記テンプレートとを比較することを特徴とする請求項２８に記載のコンピュータ読み取り可能な記録媒体。
【請求項３１】　前記テンプレートは、半構造化文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項２８に記載のコンピュータ読み取り可能な記録媒体。
【請求項３２】　前記テンプレートは、半構造化文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、
前記検索結果生成処理は、前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項２８に記載のコンピュータ読み取り可能な記録媒体。
【請求項３３】　前記テンプレートは、半構造化文書が互いに異なる要素からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項２８に記載のコンピュータ読み取り可能な記録媒体。
【請求項３４】　オープンネットワーク上の複数のサーチエンジンを介して情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、
検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を、オープンネットワーク上でのサーチエンジンの所在を示す所在情報から得る文書所在

探索処理と、
前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件が指定されたサーチエンジンを、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報から得て、検索対象サーチエンジンとして選択するサーチエンジン選択処理と、
前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定処理と、
前記変換された問い合わせ群のそれぞれを、必要に応じ、ユーザーの項目の表現形式と各ＨＴＭＬ文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換処理と、
前記変換された問い合わせを前記得られた所在に送信して、ＨＴＭＬ文書を取得する文書検索処理と、
各サーチエンジンから取得されたＨＴＭＬ文書からなる第１の検索結果から、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用いて項目を条件検索するための属性を規定する属性情報に基づいて前記抽出された項目データを選択して、第２の検索結果とする検索結果生成処理と、
前記第２の検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体。
【請求項３５】　前記検索結果生成処理は、
前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応し、ＨＴＭＬ文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とすることを特徴とする請求項３４に記載のコンピュータ読み取り可能な記録媒体。
【請求項３６】　前記検索結果生成処理は、前記検索結果を、表形式に整形することを特徴とする請求項３５に記載のコンピュータ読み取り可能な記録媒体。
【請求項３７】　前記検索結果生成処理は、前記テンプレート中の前記抽出テキスト形式情報が、他のＨＴＭＬ文書へのリンク情報を含む場合には、リンク先のＨＴＭＬ文書をスキャンして、前記リンク先のＨＴＭＬ文書と前記テンプレートとを比較することを特徴とする請求項３５に記載のコンピュータ読み取り可能な記録媒体。
【請求項３８】　前記テンプレートは、ＨＴＭＬ文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項３５に記載のコンピュータ読み取り可能な記録媒体。
【請求項３９】　前記テンプレートは、ＨＴＭＬ文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、
前記検索結果生成処理は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項３５に記載のコンピュータ読み取り可能な記録媒体。
【請求項４０】　前記テンプレートは、ＨＴＭＬ文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項３５に記載のコンピュータ読み取り可能な記録媒体。
【請求項４１】　オープンネットワーク上の任意のＨＴＭＬ文書に内在する情報から項目ごとに情報を抽出する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、
取得されたＨＴＭＬ文書に対応し、ＨＴＭＬ文書ごとに、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを解析するテンプレート解析処理と、
前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致する項目の項目データを抽出する項目データ抽出処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体。
【請求項４２】　前記項目データ抽出処理は、前記抽出された項目データを、表形式に整形することを特徴とする請求項４１に記載のコンピュータ読み取り可能な記録媒体。
【請求項４３】　前記項目データ抽出処理は、前記テンプレート中の前記抽出テキスト形式情報が、他のＨＴＭＬ文書へのリンク情報を含む場合には、リンク先のＨＴＭＬ文書をスキャンして、前記リンク先のＨＴＭＬ文書

と前記テンプレートとを比較することを特徴とする請求項４１に記載のコンピュータ読み取り可能な記録媒体。

【請求項４４】　前記テンプレートは、ＨＴＭＬ文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記項目データ抽出処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項４１に記載のコンピュータ読み取り可能な記録媒体。

【請求項４５】　前記テンプレートは、ＨＴＭＬ文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、

前記項目データ抽出処理は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項４１に記載のコンピュータ読み取り可能な記録媒体。

【請求項４６】　前記テンプレートは、ＨＴＭＬ文書が異なる項目を有する複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記項目データ抽出処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項４１に記載のコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびにそのプログラムを格納する記録媒体に関する。特に、オープンなネットワーク上に散在する複数の半構造化文書や複数のサーチエンジンが存在する環境において、これら半構造化文書の文書構造・表現形式・構成要素や、これらサーチエンジンの入力フォーム等の相違に拘わらず、各半構造化文書の所在情報・文書構造情報・項目情報・表現形式情報等を管理することによって、各半構造化文書に跨って内在する情報への、統一インターフェースによる一括したアイテムベースの統合的検索方式を実現する半構造化文書情報統合検索技術に関する。

【０００２】

【従来の技術】近年、パソコンの高機能化および低価格化、ネットワーク技術の向上、ネットワーク・プロバイダの普及と低価格化等により、インターネットに代表されるオープンネットワークが普及している。このオープンネットワークの普及に伴い、多くの情報プロバイダがコンテンツ作成を容易に実現するハイパーテキストの記述言語であるＨＴＭＬ（Hyper Text Markup Language）を用い、オープンネットワーク上で多種多様な情報発信を行うようになってきた。これら情報プロバイダは、情報コンシューマが爆発的に増加したのに伴い、急激に増加してきた。情報プロバイダが増加することにより、様々な種類の情報がネットワーク中に蓄積されてきたため、これらの情報の中から、いかに情報コンシューマが要求する情報を効率よく探索するかが大きな課題となっている。

【０００３】情報コンシューマは、これらの複数の情報源に対して興味対象の情報を一括して横断的に検索したいという要請を持つ。しかし、各ＨＴＭＬ文書の構造、表現形式、各ＨＴＭＬ文書の検索方式などは互いに異なるため、異なる情報源を横断的に検索することは困難である。

【０００４】ネットワーク上に散在するＨＴＭＬ文書の検索に関し、一般にサーチエンジンと呼ばれる情報検索装置が普及している。ここで、サーチエンジンとは、フォーム入力によりある情報を検索するシステムの総称である。図５１は、従来技術におけるＵＲＬサーチエンジンによる情報検索方式を説明する図である。ここで、ＵＲＬサーチエンジンとは、キーワードや条件入力による問い合わせに対して、ＵＲＬを検索結果として返す情報検索装置をいう。例えば「予算１０万円以下のＰＣが欲しい」という要求を満たすため、ユーザーはまずＵＲＬサーチエンジンに対してキーワード検索を行う。図５２は従来技術における各ＵＲＬサーチエンジンの構成を示す図である。ネットワークに散在するＨＴＭＬ文書検索用のキーワードと文書の所在を示すＵＲＬが予めＵＲＬサーチエンジン９００にキーワードインデックス９１０として登録されている。検索処理部９３０は、指定されたキーワードによりキーワードインデックス９１０を探索し、指定されたキーワードやその類義語を含むＨＴＭＬ文書の所在を示すＵＲＬのリストや概要を検索結果としてユーザーに返す。図５１に戻り、ユーザーは得られたＵＲＬのＨＴＭＬ文書に個別にアクセスして所望する情報を人手で探索していた。即ち、ＨＴＭＬ文書に内在する情報を検索する場合、所在が既知でないＨＴＭＬ文書から所望する情報を得るためには、ユーザーはまず全文検索によりＨＴＭＬ文書の所在を探索し、得られた所在リストの複数のＨＴＭＬ文書の内容閲覧を繰り返すため、所望する情報を取得するまでに多くの時間と労力を要していた。さらに、この所望する情報が複数のＨＴＭＬ文書に散在する場合、これらを横断的に検索することは困難であった。

【０００５】この従来の方式では、指定されたキーワードやその類義語を含むＨＴＭＬ文書の所在情報の検索はできるが、各ＨＴＭＬ文書に内在する情報に対するアイテムベースの一括した検索が不可能であった。また検索結果に対する条件指定（日付によるフィルタリング等）も不可能であった。さらに、各ＨＴＭＬ文書への検索インターフェースを入力フォームとして提供しているＵＲＬサーチエンジンを利用する場合、ユーザーがＵＲＬサ

ーチエンジンごとに個別のフォーム入力インターフェースを意識する必要があり、かつＵＲＬサーチエンジンごとに個別にアクセスする必要があった。

【０００６】具体的には、例えば、オープンネットワーク上で、オンライン商取引を行うオンラインショップのＨＴＭＬ文書は、取り扱い商品に関する情報（例えば、商品名や価格など）を１つの意味のあるデータ群として、表形式や箇条書きの形式でリスト記述する場合が非常に多い。これらオンラインショップのＨＴＭＬ文書に内在する情報を横断的に検索することに対する需要が高まっている。この横断的検索として、例えば”指定の商品を最安値で販売しているショップの検索”等がある。従来これらオンラインショップのＨＴＭＬ文書から横断的検索を行うためには、ユーザは欲しい商品の名前、メーカ名、商品種別等をキーワードとして、図５１の検索方式によりＨＴＭＬ文書の所在情報を取得し、得られた所在のＨＴＭＬ文書に１つずつアクセスし、所望の商品の有無を確認する必要があった。しかし図５１の検索方式は、文書の構成要素を考慮しない全文検索であるため、全く関係のないＨＴＭＬ文書の所在まで大量に検索してしまい、これら大量のＨＴＭＬ文書の中から人手で所望の商品情報を探索するのに多くの時間と労力を要していた。

【０００７】このように従来の検索方式では、ＨＴＭＬ文書中の情報を項目別に収集することができなかった。即ち”商品情報を記載した表を内包するＨＴＭＬ文書”に対しては”商品価格””商品イメージ””メーカ名”等、”店舗情報が箇条書きで記載されているＨＴＭＬ文書”に対しては”店舗名””電話番号””住所”等の項目別に情報を抽出することが困難であった。また、ＨＴＭＬ文書からの検索結果に対して日付によるフィルタリング処理などの条件指定を行うこともできなかった。

【０００８】これら項目別に管理されている情報を抽出するために、文書内部の構造や文書間の関連を独自のモデルにマッピングすることにより、仮想的なデータベースを作成する従来技術がある。この従来技術の１つの例は、N.Ashish, C.A.Knoblock, "Semi-automatic wrapper generation for internet information sources",Proceedings of cooperative information systems,1997.に開示されている。この技術は、ＨＴＭＬ文書中で特定のタグ（ＴＩＴＬＥタグ、Ｈ１タグ等）や、特定のフォントタグの内容（大きさ、色、太字・イタリック等の書体等）を持つ箇所を意味のある情報ととらえ、これらの情報を自動的に抽出するための技術である。この技術は、１つの情報の最小のまとまりが１つのＨＴＭＬ文書に記述され、これらＨＴＭＬ文書が同じ形式で記述された複数のＨＴＭＬ文書を対象としている。この技術は、例えば、地域ごとの気象情報が異なるＨＴＭＬ文書に記述されている場合に有効である。

【０００９】しかし、この技術は、１つのＨＴＭＬ文書に表形式や箇条書きの形式でデータ群をリスト記述することは考慮されていないため、上記のケースには適用できない。

【００１０】従来技術の他の例は、J.Hammer, H.Garcia-Molina, J.Cho, R.Araha, A.Crespo, "Extracting semistructured information from the web", Workshop on management of semistructured data,1997.に開示されている。この技術は、ＯＥＭという独自のデータモデルを用いて下層のデータベースを構築し、このデータベースと様々な情報源の対応を管理することにより、複数の異種情報源の統合的な検索を実現する技術である。この対応管理のため、この技術はＨＴＭＬ文書に対してはＨＴＭＬタグ記述に依存したテンプレートファイルを用いる方式を採用している。

【００１１】しかし、この技術は、ＨＴＭＬ文書に変更が生じると仮想のデータベースに影響が及び、仮想のデータベースに変更が生じるとアプリケーションに影響が及ぶため、システムの運用、保守に多大な労力が必要であった。

【００１２】さらに、オンラインショップの取扱商品情報等のためのＨＴＭＬ記述には、標準化された形式がないため、各ＨＴＭＬ文書に以下の差異が生じている。

【００１３】第１に、ショップにより各ＨＴＭＬ文書の文書構造が異なる。例えば、ショップＡの取扱商品はＴＡＢＬＥタグで記述される表形式で提示されたり、ショップＢの取扱商品はＵＬタグで記述される箇条書きで提示されたりしている。

【００１４】第２に、ＨＴＭＬ文書上の同一の取扱商品に関する情報の表現形式が異なる。例えば、価格を表す表現形式では、円、千円、万円、＄等の単位の違いや、全角、半角等の表記の違いがある。

【００１５】第３に、ＨＴＭＬ文書の同じ情報を表すデータの構成要素が異なる。例えば、商品名を示すデータの構成要素は、商品名のみの記述、商品名と型番を併せて記述、メーカ名と商品名と型番を併せて記述、等の違いがある。従来の検索方式で取得したＨＴＭＬ文書から所望の情報を得るため、ユーザは、これらの相違する情報を並べて比較する必要がある。これらの情報の中から所望の商品情報を探索するのにも、多くの時間と労力を要していた。

【００１６】さらに、複数のサーチエンジンを用いてオープンネットワーク上の情報を検索する場合、これらのサーチエンジンにはそれぞれ取り扱う情報の種類等の差異があるため、状況に応じてユーザが使い分ける必要がある。換言するとユーザーは各サーチエンジンに検索要求を発行する際に、サーチエンジン固有の所在情報、検索インタフェースを意識する必要があった。

【００１７】このため、第１に、ユーザによるサーチエンジンの所在情報の管理が困難であった。ユーザはサーチエンジンの所在情報を、ブックマーク等を用い個人で

管理する必要があるため、特にモバイル環境下など自端末以外の環境での検索が困難であった。

【0018】第2に、各種サーチエンジンの普及に伴う入力フォームの提供する検索インタフェースの不統一性が生じた。各サーチエンジンの普及に伴い、入力フォームが乱立している。しかし、入力フォームの構造は統一されていないため、ユーザーはサーチエンジン毎に固有の操作体系、操作手順を把握する必要がある。またユーザーは、ある検索項目の処理にどのサーチエンジンが有効であるかを把握することができない。かつ得られたHTML文書中の情報を条件処理することができない。

【0019】第3に、サーチエンジンへの検索効率の悪化が生じた。上述したようにユーザーは、所望の情報を得るまで人手でサーチエンジン毎に検索を行うため、検索回数が増加し、効率が非常に悪い。

【0020】第4に、各サーチエンジンからの検索結果の項目、表現形式、文字コードなどのフォーマットが不統一であるため、ユーザーが検索結果を比較するのが困難である。

【0021】上記の各サーチエンジンの異種性を解消するため、サーチエンジンの一種であるURLサーチエンジンの共通な検索インタフェースを作成し、当該検索インタフェースと個々のURLサーチエンジンの検索インタフェースの対応を管理し、共通検索インタフェースに対する検索要求を個々の検索エンジンの検索要求に変換/実行する従来技術が、Jumon World Seek,"http://member.nifty.ne.jp/jumon "に開示されている。この技術は、共通検索インタフェースがテキストボックス1つから構成されるURLサーチエンジンを提供する。しかし、一般にURLサーチエンジンだけではなく多種多様なサーチエンジンが存在し、これらの横断的な検索を実現するためには、以下の問題点があった。

【0022】(1)複数の入力項目の考慮が必要。

【0023】最もシンプルな入力フォーム構成では、入力項目は検索する用語を入力するテキストボックス(キーワード入力部)のみであるが、キーワードとともに他の検索条件(エリア、業種等)について同時に入力し、絞り込み検索を行うことについて配慮されている場合もある。この場合、HTMLは項目を有さない半構造化文書であるため、従来技術はシステムとして複数の入力項目をサポートできず、絞り込み検索はできなかった。

【0024】(2)使用される入力フォームの多様性への対応が必要。

【0025】サーチエンジンで通常用いられるテキスト入力用の入力フォームのオブジェクトには、テキストボックス、複数項目中1項目を選択するラジオボタン、複数項目中で任意の複数項目を選択するセレクトボックスまたはチェックボックスなど要求条件を適切に入力するためのオブジェクトが複数存在する場合がある。この場合、従来技術ではシステムとしてテキストボックス以外のオブジェクトをサポートしていないため、対応することができなかった。

【0026】(3)さらに、複数のサーチエンジンにわたる共通検索インタフェースを用いる場合、この共通検索インターフェースの修正時にアプリケーションを再構築する必要がある。

【0027】共通検索インターフェースに対してサーチエンジンの追加/修正/削除を行う際に、共通検索インタフェースの修正が必要になり、対応するアプリケーションを再構築しなければならない。

【0028】すなわち、従来技術においては、多種多様なサーチエンジンを取り込むことができず、システム構築/維持管理に多くの時間と労力が必要であった。

【0029】

【発明が解決しようとする課題】本発明は、上記の問題点を解決するためになされたものである。

【0030】そして、その目的とするところは、オープンなネットワークに散在する複数のHTML文書に内在する情報の文書構造、表現形式、構成要素などが互いに異なっていても、これら文書を跨った情報検索を実現し、このHTML記述上の差異をユーザーごとの統一形式に変換した検索結果を返却することのできる、半構造化文書情報統合検索体系を提供することにある。

【0031】本発明の他の目的は、オープンなネットワークに複数のサーチエンジンが存在する環境において各サーチエンジン固有の入力フォームのオブジェクトを個別に管理することにより複数のサーチエンジンの異種性を解消し、ユーザーの検索要求に対して各サーチエンジン固有の検索要求を生成して検索を実行することのできる、半構造化文書情報統合検索体系を提供することにある。

【0032】本発明の他の目的は、HTML文書の所在情報、HTML文書に内在する文書の構造情報、各構成要素の属性情報をHTML文書ごとに管理することにより、所在、文書構造、属性が互いに異なる任意のHTML文書から情報を項目別に抽出することのできる半構造化文書情報統合検索体系を提供することにある。

【0033】

【課題を解決するための手段】本発明の特徴は、オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索装置であって、半構造化文書ごとに該半構造化文書中から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報を記憶する記憶部と、入力された問い合わせから、前記メタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得る検索部と、ユーザーごとに所定の単一フォーマットで前記検索結果を出力する出力部とを具備することを特徴とする半構造化文書情報統合検索装置を提供する点にある。

【0034】また、本発明の他の特徴は、オープンネッ

トワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索装置であって、オープンネットワーク上での半構造化文書の所在を示す所在情報と、前記半構造化文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報と、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶する記憶部と、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を前記所在情報から得る文書所在探索部と、入力された前記問い合わせを、必要に応じ、前記表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換する問い合わせ変換部と、前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得する文書検索部と、取得された各半構造化文書から、前記文書構造情報に基づいて、項目データを抽出し、必要に応じて前記検索条件を用い、前記属性情報に基づいて前記抽出された項目データを選択して検索結果とする文書処理部と、前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部とを具備することを特徴とする半構造化文書情報統合検索装置が提供される。

【００３５】また、本発明の他の特徴によれば、上記半構造化文書情報統合検索装置は、さらに、半構造化文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、半構造化文書から抽出すべき項目群の所定の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部を具備し、前記文書処理部は、前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応する前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とする点にある。

【００３６】また、本発明の他の特徴は、前記文書処理部は、前記検索結果を、表形式に整形する点にある。

【００３７】また、本発明の他の特徴は、前記文書処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他の半構造化文書へのリンク情報を含む場合には、リンク先の半構造化文書をスキャンして、前記リンク先の半構造化文書と前記テンプレートとを比較する点にある。

【００３８】また、本発明の他の特徴は、前記テンプレートは、半構造化文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００３９】また、本発明の他の特徴は、前記テンプレートは、半構造化文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記文書処理部は、前記取得された半構造化文書をスキャンして、該半構造化文書の前記部分構造と、該半構造化文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【００４０】また、本発明の他の特徴は、前記テンプレートは、半構造化文書が互いに異なる要素からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００４１】また、本発明の他の特徴は、オープンネットワーク上の複数のサーチエンジンを介して情報を検索する半構造化文書情報統合検索装置であって、オープンネットワーク上でのサーチエンジンの所在を示す所在情報と、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報と、ＨＴＭＬ文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報と、ユーザーの項目の表現形式と各ＨＴＭＬ文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶する記憶部と、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を前記所在情報から得る文書所在探索部と、前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件を満たす検索条件が指定されたサーチエンジンを、前記入力必須項目情報から得て、検索対象サーチエンジンとして選択するサーチエンジン選択部と、前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定部と、前記変換された問い合わせ群のそれぞれを、必要に応じ、前記表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換部と、前記変換された問い合わせを前記得られた所在に送信して、ＨＴＭＬ文書を取得する文書検索部と、各サーチエンジンから取得されたＨＴＭＬ文書からなる第１の検索結果から、前記文書構造情報に基づいて、項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用い、前記属性情報に基づいて、前記抽出された項目データを選択して、第２の検索結果

とする検索結果処理部と、前記第２の検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部とを具備することを特徴とする半構造化文書情報統合検索装置を提供する点にある。

【００４２】また、本発明の他の特徴は、上記半構造化文書情報統合検索装置は、さらに、ＨＴＭＬ文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部を具備し、前記文書処理部は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、第２の検索結果とする点にある。

【００４３】また、本発明の他の特徴は、前記文書処理部は、前記検索結果を、表形式に整形する点にある。

【００４４】また、本発明の他の特徴は、前記文書処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他のＨＴＭＬ文書へのリンク情報を含む場合には、リンク先のＨＴＭＬ文書をスキャンして、前記リンク先のＨＴＭＬ文書と前記テンプレートとを比較する点にある。

【００４５】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００４６】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記文書処理部は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【００４７】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００４８】また、本発明の他の特徴は、オープンネットワーク上の任意のＨＴＭＬ文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出装置であって、ＨＴＭＬ文書ごとに、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部と、取得されたＨＴＭＬ文書に対応するテンプレートを解析するテンプレート解析部と、前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理部とを具備することを特徴とする半構造化文書情報抽出装置を提供する点にある。

【００４９】また、本発明の他の特徴は、前記テンプレート処理部は、前記抽出された項目データを、表形式に整形する点にある。

【００５０】また、本発明の他の特徴は、前記テンプレート処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他のＨＴＭＬ文書へのリンク情報を含む場合には、リンク先のＨＴＭＬ文書をスキャンして、前記リンク先のＨＴＭＬ文書と前記テンプレートとを比較する点にある。

【００５１】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００５２】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記テンプレート処理部は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【００５３】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００５４】また、本発明の他の特徴は、オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索方法であって、入力された問い合わせから、半構造化文書ごとに該半構造化文書から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得るステップと、ユーザーごとに所定の単一フォーマットで前記検索結果を出力するステップとを含むことを特徴とする半構造化文書情報統合検索方法を提供する点にある。

【００５５】また、本発明の他の特徴は、オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索方法であって、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を、オープンネットワーク上での半構造化文書の所在を示す所在情報から得るステップと、入力された前記問い合わせを、必要に応じ、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換するステップと、前記変換された検索要求を前記得られた所在に送信して、半構造化文書を取得するステップと、取得された各半構造化文書から、半構造化文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて前記検索条件を用い、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報に基づいて、前記抽出された項目データを選択して検索結果とするステップと、前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換するステップとを含むことを特徴とする半構造化文書情報統合検索方法を提供する点にある。

【００５６】また、本発明の他の特徴は、オープンネットワーク上の複数のサーチエンジンを介して情報を検索する半構造化文書情報統合検索方法であって、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を、オープンネットワーク上でのサーチエンジンの所在を示す所在情報から得るステップと、前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件が指定されたサーチエンジンを、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報から得て、検索対象サーチエンジンとして選択するステップと、前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換するステップと、前記変換された問い合わせ群のそれぞれを、必要に応じ、ユーザーの項目の表現形式と各ＨＴＭＬ文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換するステップと、前記変換された問い合わせを前記得られた所在に送信して、ＨＴＭＬ文書を取得するステップと、各サーチエンジンから取得されたＨＴＭＬ文書からなる第１の検索結果から、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用いて項目を条件検索するための属性を規定する属性情報に基づいて前記抽出された項目データを選択して、第２の検索結果とするステップと、前記第２の検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換するステップとを含むことを特徴とする半構造化情報統合検索方法を提供する点にある。

【００５７】また、本発明の他の特徴は、オープンネットワーク上の任意のＨＴＭＬ文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出方法であって、取得されたＨＴＭＬ文書に対応し、ＨＴＭＬ文書ごとに、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを解析するステップと、前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するステップとを含むことを特徴とする半構造化文書情報抽出方法を提供する点にある。

【００５８】また、本発明の他の特徴は、オープンネットワーク上の複数の半構造化文書に内在する情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、入力された問い合わせから、半構造化文書ごとに該半構造化文書から抽出すべき項目および該項目を条件検索するための項目情報を定義するメタ情報に基づいて、複数の半構造化文書に散在する情報を検索して一括した検索結果を得る処理と、ユーザーごとに所定の単一フォーマットで前記検索結果を出力する処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体を提供する点にある。

【００５９】また、本発明の他の特徴は、オープンネットワーク上の複数の半構造化文書に内在する情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を、オープンネットワーク上での半構造化文書の所在を示す所在情報から得る文書所在探索処理と、入力された前記問い合わせを、必要に応じ、ユーザーの項目の表現形式と各半構造化文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記得られた所在の半構造化文書中の前記検索項目に対応する項目の表現形式に変換する問い合わせ変換処理と、前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得する文書検索処理と、

取得された各半構造化文書から、半構造化文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて前記検索条件を用い、前記項目ごとに前記項目を条件検索するための属性を規定する属性情報に基づいて、前記抽出された項目データを選択して検索結果とする検索結果生成処理と、前記検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体を提供する点にある。

【００６０】また、本発明の他の特徴は、前記検索結果生成処理は、前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応し、半構造化文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、半構造化文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とする点にある。

【００６１】また、本発明の他の特徴は、前記検索結果生成処理は、前記検索結果を、表形式に整形する点にある。

【００６２】また、本発明の他の特徴は、前記検索結果生成処理は、前記テンプレート中の前記抽出テキスト形式情報が、他の半構造化文書へのリンク情報を含む場合には、リンク先の半構造化文書をスキャンして、前記リンク先の半構造化文書と前記テンプレートとを比較する点にある。

【００６３】また、本発明の他の特徴は、前記テンプレートは、半構造化文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００６４】また、本発明の他の特徴は、前記テンプレートは、半構造化文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記検索結果生成処理は、前記取得された半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【００６５】また、本発明の他の特徴は、前記テンプレートは、半構造化文書が互いに異なる要素からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００６６】また、本発明の他の特徴は、オープンネットワーク上の複数のサーチエンジンを介して情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を、オープンネットワーク上でのサーチエンジンの所在を示す所在情報から得る文書所在探索処理と、前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索条件が指定されたサーチエンジンを、各サーチエンジンへの入力フォームに対する入力必須項目を定義する入力必須項目情報から得て、検索対象サーチエンジンとして選択するサーチエンジン選択処理と、前記検索項目および前記検索条件と、各サーチエンジンの有する項目および前記入力必須項目との組み合わせを規定するマトリックステーブルに基づき、各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定処理と、前記変換された問い合わせ群のそれぞれを、必要に応じ、ユーザーの項目の表現形式と各ＨＴＭＬ文書の項目の表現形式との間の変換情報を定義する表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換処理と、前記変換された問い合わせを前記得られた所在に送信して、ＨＴＭＬ文書を取得する文書検索処理と、各サーチエンジンから取得されたＨＴＭＬ文書からなる第１の検索結果から、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づいて項目データを抽出し、必要に応じて対応する前記検索処理パターンに従い、前記検索条件を用いて項目を条件検索するための属性を規定する属性情報に基づいて前記抽出された項目データを選択して、第２の検索結果とする検索結果生成処理と、前記第２の検索結果を、必要に応じ、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体を提供する点にある。

【００６７】また、本発明の他の特徴は、前記検索結果生成処理は、前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応し、ＨＴＭＬ文書ごとに、前記文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出して、検索結果とする点にある。

【００６８】また、本発明の他の特徴は、前記検索結果生成処理は、前記検索結果を、表形式に整形する点にある。

【００６９】また、本発明の他の特徴は、前記検索結果生成処理は、前記テンプレート中の前記抽出テキスト形

式情報が、他のＨＴＭＬ文書へのリンク情報を含む場合には、リンク先のＨＴＭＬ文書をスキャンして、前記リンク先のＨＴＭＬ文書と前記テンプレートとを比較する点にある。

【００７０】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００７１】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記検索結果生成処理は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【００７２】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記検索結果生成処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００７３】また、本発明の他の特徴は、オープンネットワーク上の任意のＨＴＭＬ文書に内在する情報から項目ごとに情報を抽出する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、取得されたＨＴＭＬ文書に対応し、ＨＴＭＬ文書ごとに、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを解析するテンプレート解析処理と、前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致する項目の項目データを抽出する項目データ抽出処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体を提供する点にある。

【００７４】また、本発明の他の特徴は、前記項目データ抽出処理は、前記抽出された項目データを、表形式に整形する点にある。

【００７５】また、本発明の他の特徴は、前記項目データ抽出処理は、前記テンプレート中の前記抽出テキスト形式情報が、他のＨＴＭＬ文書へのリンク情報を含む場合には、リンク先のＨＴＭＬ文書をスキャンして、前記リンク先のＨＴＭＬ文書と前記テンプレートとを比較する点にある。

【００７６】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記項目データ抽出処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００７７】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書の各部分構造に対して、前記部分構造の一部に存在する不均一な部分を透過に取得するための複数の抽出テキスト形式情報が記述され、前記項目データ抽出処理は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出する点にある。

【００７８】また、本発明の他の特徴は、前記テンプレートは、ＨＴＭＬ文書が異なる項目を有する複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、前記項目データ抽出処理は、抽出された項目データを、各部分構造ごとの検索結果とする点にある。

【００７９】

【発明の実施の形態】以下において、図面を用いて本発明に係る半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびにそのプログラムを格納する記録媒体の実施形態が詳細に説明される。尚、以下の実施形態において、半構造化文書とは、ＨＴＭＬ（Hyper Text Markup Language）文書、ＳＧＭＬ（StandardGeneralized Markup Language）文書、ＸＭＬ（eXtensive Markup Language）文書を含む。以下、特に記載しない限り、半構造化文書をＨＴＭＬ文書で代表させて説明する。尚、以下の実施形態は、ＳＧＭＬ文書およびＸＭＬ文書に対しても、適宜修正して適用することができる。例えば情報検索用のサーチエンジンが具備する入力フォームなどもこのＨＴＭＬ文書により構成されており、以下、ＨＴＭＬ文書にはこれら入力フォームを含むものとする。また、以下の実施形態は、例えばエレクトロニック・コマース、電子図書館や電子カタログからの情報検索など、オープンネットワーク上の複数の互いに種々の相違を有する複数のＨＴＭＬ文書を利用するアプリケーションに広範に適用しうる。

【００８０】まず、図１および図２を参照して、本発明に係る半構造化文書情報統合検索体系の原理を説明する。

【００８１】図１は、本発明に係る半構造化文書情報統合検索体系を用いる場合のユーザーの処理手順のイメージを示す。本発明に係る装置は、ユーザーから入力される検索要求（例えば、「１０万円以下のＰＣが欲しい」）に基づいて、複数のＨＴＭＬ文書をユーザーに替わって柔軟に検索し、検索処理結果を一括してユーザーに送信する。この検索要求は、従来の検索用キーワードではなく、検索項目とその条件からなる簡易な構文の問い合わせ文を用いることができる。即ち、「１０万円以

下」などの条件を含む検索を処理することができる。

【００８２】ＨＴＭＬ文書は、ＲＤＢのような項目単位で構造化されたデータと異なり、基本的にはプレーンテキストであるが、タグなどによりある程度データが構造化されている、いわゆる半構造化データであることを特徴とする。例えばＨＴＭＬ文書に内在する表、リスト、箇条書きなどの意味のある１つのデータ群が、複数のＨＴＭＬ文書を跨って保有されていたり、逆に複数のデータ群が１つのＨＴＭＬ文書に内在される場合がある。同時にこれらのデータ群のうち、ある項目に対応するデータを条件処理することができない。また、ＨＴＭＬにより記述された検索用の入力フォームを有するサーチエンジンでは、検索条件として指定可能なデータ項目が固定であったり、検索条件として必須のデータ項目があったりする。こうした構造を有するＨＴＭＬ文書に対して、本発明に係る装置は柔軟な検索要求に対応する検索結果を一括して出力する。

【００８３】図２は、本発明に係る装置の原理を示す。本発明に係る装置は、事前に登録された、各ＨＴＭＬ文書ごとの、ＨＴＭＬ文書の所在情報、文書構造情報、表現形式などを、ＨＴＭＬ文書メタデータとしてＨＴＭＬ文書メタデータ記憶部１５中に管理する。この所在情報は、例えばＵＲＬである。この文書構造情報は、ＨＴＭＬ文書中の表、リスト、箇条書きなどの部分構造に関する情報であり、これらのデータを抽出すべき各項目にマッピングするための情報である。具体的にはこの文書構造情報は、抽出すべき項目に対応するデータがタグやスラッシュなどのデリミタで区切られているという情報であり、タグやスラッシュなどのデリミタにより識別されるＨＴＭＬ文書データの区切られた領域を、項目に対応付けて表形式で管理される。この表現形式は、同じ意味を表すデータが異なる表現でＨＴＭＬ文書内に存在する場合の、それぞれの表現形式である。

【００８４】ユーザーは、本装置の問い合わせ処理部１３に、問い合わせを入力する。問い合わせ処理部１３は、ＨＴＭＬ文書メタデータ記憶部１５に事前に登録されている情報を参照して、ＨＴＭＬ文書の所在、構造、表現形式を特定する。問い合わせ処理部１３は、各ＨＴＭＬ文書を取得し、特定されたこれらの情報を用いて、各ＨＴＭＬ文書に内在する情報を抽出し、必要に応じて条件処理を行う。このため、本発明に係る装置は、各ＨＴＭＬ文書に内在する情報の条件検索結果を、一括してかつユーザーの表現形式に適合させてユーザーに出力することができる。従って、ユーザーは、１回の検索要求で所望する情報をネットワーク上に散在するＨＴＭＬ文書から一括して得ることができる。このため、検索効率が向上すると同時にネットワーク上のトラフィックが低減される。

【００８５】すなわち、本発明に係る装置は、第１に、オープンネットワークに接続されているＨＴＭＬ等の半構造化文書の文書構造情報を管理することにより、複数のＨＴＭＬ文書からの項目ベースでの検索を実現する。第２に、各サーチエンジンなどのＨＴＭＬ文書に散在する情報を、個々のＷｅｂ情報源に対する検索方式をユーザーに意識させずに統合的な検索を実現する。

【００８６】第１の実施形態図３から図１４を参照して、本発明に係る半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびに半構造化文書情報統合検索プログラムおよび半構造化文書情報抽出プログラムを格納した記録媒体の第１の実施形態であるＨＴＭＬ文書情報統合検索装置を詳細に説明する。

【００８７】第１の実施形態は、オープンなネットワークに散在する複数ＨＴＭＬ文書に対し、各ＨＴＭＬ文書が内在する情報の文書構造、表現形式、表などの部分構造の構成要素が互いに異なっていても、各ＨＴＭＬ文書を跨った情報検索を実現し、またそのＨＴＭＬ記述上の差異をユーザ毎の統一形式に変換して検索結果を一括して返却するものである。

【００８８】第１の実施形態の構成の詳細な説明を行う前に、第１の実施形態で用いられる表現形式の概念および関連する用語について説明する。ＨＴＭＬ文書が異なると、同じ意味を表す情報でも、異なった表現形式を用いていることがある。例えば、製品価格に対して、「￥１，０００」、「一千円」、「１０００円」と、ＨＴＭＬ文書の記述者によって、様々な表現形式が存在する。そこで、以下の通り、用語を定義する。

【００８９】・ドメインとは、１つの表現形式である。

【００９０】例）価格で、１,000円は、「円」つき表現形式で１つのドメイン。

【００９１】価格で、￥1,000 は、「￥」つき表現形式で１つのドメイン。

【００９２】・ドメイングループとは、同じ意味を表すドメインの集まりである。

【００９３】例）価格、年月日、等。

【００９４】・ユーザ入力ドメインとは、ユーザ毎の検索要求の入力に用いるドメインである。

【００９５】例）価格は、「円」つき表現形式、年月日は西暦かつ「／」区切り表現形式。

【００９６】・ユーザ出力ドメインとは、ユーザ毎に指定する検索結果に用いるドメインである。

【００９７】例）価格は、「￥」つき表現形式、年月日は年号略号かつ「．」区切り表現形式。

【００９８】・ユーザドメインとは、ユーザ入力ドメインとユーザ出力ドメインの総称である。

【００９９】・ローカルドメインとは、各ＨＴＭＬ文書上のドメインである。

【０１００】例）価格は、「円」つき表現形式。

【０１０１】・ドメイン変換関数とは、ユーザ入力ドメインからローカルドメイン、ローカルドメインからユーザ出力ドメインへの変換を行う関数である。

【0102】なお、ユーザ入力ドメイン、ユーザ出力ドメイン、ローカルドメインが相互に異なる場合、これらの差異はドメイン変換関数を用いて解消される。

【0103】図3は、第1の実施形態に係るHTML文書情報統合検索装置の構成を示すブロック図である。HTML文書情報統合検索装置1は、ユーザーインターフェース部11と、構文解析部12と、問い合わせ処理部13と、HTML文書アクセス部14と、HTML文書メタデータ記憶部15と、HTML文書メタデータ管理部16とにより構成される。問い合わせ処理部13は、問い合わせ項目探索部131と、問い合わせ変換部132と、変換関数ライブラリ133と、HTML文書処理部134と、検索結果変換部135とを有する。

【0104】ユーザインタフェース部11は、ユーザのアプリケーションプログラム3から入力される検索項目と検索条件からなる問い合わせ文を受け付ける。構文解析部12は、ユーザインタフェース部11で受け付けた問い合わせ文の構文を解析する。問い合わせ処理部13は、各HTML文書に内在する情報から所望する項目情報を一括して検索する。問い合わせ処理部13中で、問い合わせ項目探索部131は、問い合わせ文中で指定された項目の所在を探索する。問い合わせ変換部132は、問い合わせ文のユーザ入力ドメインをローカルドメインに変換し、HTML文書アクセス部14が送出する問い合わせ文を生成する。HTML文書処理部134は、HTML文書アクセス部14が受信したHTML文書から取得した情報に対し、問い合わせ文に応じた処理（検索項目の選択、検索条件によるフィルタリング等）を行う。検索結果変換部135は、検索されたデータの表現形式をローカルドメインからユーザ出力ドメインに変換する。HTML文書アクセス部14は、オープンなネットワークに散在するHTML文書を取得し、その内在する情報を例えば表形式などの統一フォーマットに変換する。このHTML文書アクセス部14には、HTML文書21およびこのHTML文書21を管理するWEBサーバ22からなる複数のHTML文書サーバ2-1，2-2・・・が接続されている。HTML文書メタデータ記憶部15は、各HTML文書の文書構造、HTML文書に内在する情報の表現形式や項目情報などの検索対象となるHTML文書に関する情報を記憶し管理する。この項目情報として、表などの部分構造中の構成要素と抽出すべき項目が1対1対応でない場合には、抽出すべき項目に対応させて部分構造中の構成要素は複数の構成要素として管理される。HTML文書メタデータ管理部16は、HTML文書メタデータ記憶部15に対する各種情報の入力／削除／変更を行う。システム管理者は、例えばエディタなどに実装されるHTML文書メタデータ管理部16を介して、HTML文書メタデータを登録・管理する。

【0105】図4は、HTML文書メタデータ記憶部15が保有するテーブルの詳細を示す。HTML文書メタデータは、各HTML文書の所在情報を管理するHTML文書テーブル151と、各HTML文書に内在する構成要素の表形式に変換するための情報を管理するHTML文書－表マッピングテーブル152と、各HTML文書の項目ごとにこの項目の属性を管理するHTML文書項目テーブル153と、各ドメインの表現形式を管理するドメインテーブル154と、ユーザーごとに入力ドメインおよび出力ドメインを管理するユーザドメインテーブル155と、ドメイン変換関数を管理するドメイン変換関数テーブル156とにより構成される。

【0106】次に、第1の実施形態に係るHTML文書情報統合検索装置1の処理手順の概略を説明する。第1の実施形態の処理手順は、図5に示す検索を実行する前にHTML文書メタデータ管理部16を介して管理者がメタデータの準備を行う準備フェーズと、図6に示す検索を実行する検索フェーズの2段階のフェーズがある。

【0107】図5に示す準備フェーズでは、HTML文書の所在情報をHTML文書テーブル151に定義し（ステップS100）、HTML文書に内在する情報の表形式への対応情報をHTML文書－表マッピングテーブル152に定義し（ステップS110）、HTML文書に内在する情報の項目の属性をHTML文書項目テーブル153に定義し（ステップS120）、HTML文書に内在する情報の項目のローカルドメインをドメインテーブル154に定義し（ステップS130）、ユーザ入力ドメイン、ユーザ出力ドメインをユーザードメインテーブル155に定義し（ステップS140）、必要なドメイン間の変換関数が既存か否かについてを判定し（ステップS145）、必要なドメイン変換関数が存在しない場合、新たなドメイン変換関数を作成してドメイン変換関数テーブル156に定義する（ステップS150）。

【0108】図6に示す検索フェーズでは、まず構文解析部12はユーザからの問い合わせ文を解析し、問い合わせ項目探索部131は指定された項目の所在をHTML文書テーブル151から探索し（ステップS200）、すべての検索項目に対応する項目を保持するHTML文書をHTML文書属性テーブル153から探索し（ステップS210）、問い合わせ変換部132はステップS210で探索された項目に対応するユーザ入力ドメインとユーザ出力ドメインとローカルドメインをドメインテーブル154，ユーザードメインテーブル155から取得し（ステップS220）、全項目のユーザ入力ドメインとローカルドメインが同一か否かを判定し（ステップS225）、ユーザ入力ドメインとローカルドメインとが異なる項目に対応するドメイン変換関数を取得し、当該項目をローカルドメインの表現形式に変換する（ステップS230）。HTML文書処理部134は、HTML文書アクセス部14を介して各HTML文書を

取得して検索項目に対応する項目を抽出して検索結果を得（ステップＳ２４０）、全項目のユーザ出力ドメインとローカルドメインが同一か否かを判定し（ステップＳ２４５）、検索結果変換部１３５はユーザ出力ドメインとローカルドメインが異なる項目に対し、ドメイン変換関数テーブル１５６からドメイン変換関数を取得して当該項目をユーザ出力ドメインに変換し（ステップＳ２５０）、ユーザーインターフェース部１１を介して変換後の検索結果を出力する（ステップＳ２６０）。

【０１０９】以下、第１の実施形態の処理手順の詳細を、図７から図１４を用いて具体的に説明する。

【０１１０】図７（Ａ）はショップＡの商品情報を示すＨＴＭＬ文書のＷＥＢブラウザでの表示例であり、図８（Ａ）はショップＢの商品情報を示すＨＴＭＬ文書の表示例であるとする。図７（Ｂ）は図７（Ａ）の情報を表示するためのＨＴＭＬ記述例であり、図１０（Ｂ）は図１０（Ａ）の情報を表示するためのＨＴＭＬ記述例である。

【０１１１】ＨＴＭＬ文書の構造を示す文書構造としては、ショップＡの商品情報は内在情報の表示に表形式を用いるため、ＴＡＢＬＥタグを使用している。ショップＢの商品情報は内在情報の表示に箇条書きを用いるため、ＯＬタグを使用している。

【０１１２】表現形式としては、ショップＡの商品情報は価格情報として"￥"つき表現形式を使用している。ショップＢの商品情報は価格情報として"円"つき表現形式を使用している。

【０１１３】各ＨＴＭＬ文書の有する構成要素としては、ショップＡの商品情報は商品名を、メーカ名と商品名の構成要素から構成している。ショップＢの商品情報はメーカ名と商品名が分離されて構成されている。

【０１１４】所在情報としては、ショップＡの商品情報のＵＲＬは、"http://www.shop－a.co.jp/products.html"である。ショップＢの商品情報のＵＲＬは、"http://www.shop－b.co.jp/shouhin.html"である。

【０１１５】上記の通り、図７（Ａ）の商品情報と図７（Ｂ）の商品情報とは、文書構造、表現形式、構成要素のすべての面で異なっている。

【０１１６】（１）準備フェーズ

まず、各ＨＴＭＬ文書の所在情報として、図９に示すように以下のページ名とＵＲＬをＨＴＭＬ文書テーブル１５１に設定する（図５のステップＳ１００）。

【０１１７】(a) ショップＡの商品情報

・ページ名：Shop－Ａ

・ＵＲＬ： http://www.shop－a.co.jp/products.html

(b) ショップＢの商品情報

・ページ名：Shop－Ｂ

・ＵＲＬ： http://www.shop－b.co.jp/shouhin.html

各ＨＴＭＬ文書に内在する構成要素の表形式への対応情報として、図１０に示すように以下のページ名、レコード開始点、列１～列４の取り出し方をＨＴＭＬ文書－表マッピングテーブル１５２に設定する（ステップＳ１１０）。尚、ショップＢの価格情報は、数字と"，"の箇所のみが取得されるよう設定している。

【０１１８】(a) ショップＡの商品情報

・ページ名：Shop－Ａ

・レコード開始："　<TR><TD>"で始まる行

・列１："ショップＡ"固定値

・列２：レコード開始行中の１つ目の"<TD>"と１つ目の"/"の間

・列３：レコード開始行中の１つ目の"/"と１つ目の"</TD>"の間

・列４：レコード開始行中の２つ目の"<TD>"と２つ目の"</TD>"の間

(b) ショップＢの商品情報

・ページ名：Shop－Ｂ

・レコード開始："　<LI>"で始まる行

・列１："ショップＢ"固定値

・列２：レコード開始行中の１つ目の"<LI>"と１つ目の"/"の間

・列３：レコード開始行中の１つ目の"/"と２つ目の"/"の間

・列４：レコード開始行中の２つ目の"/"と１つ目の"円"の間

次に、ＨＴＭＬ文書に内在する項目の属性情報として、図１１に示すように、以下のページ名、対応列、列名、データ型をＨＴＭＬ文書項目テーブル１５３に設定する（ステップ１２０）。ここでは、価格情報のみが数値のデータ型として定義されている。このデータ型は、検索条件の処理時に数値として比較を行うために設定されている。

【０１１９】(a-1) ページ名Ａの列１

・ページ名：Shop－Ａ

・対応列：列１

・列名：ショップ名

・データ型：文字列

(a-2) ページ名Ａの列２

・ページ名：Shop－Ａ

・対応列：列２

・列名：メーカ名

・データ型：文字列

(a-3) ページ名Ａの列３

・ページ名：Shop－Ａ

・対応列：列３

・列名：商品名

・データ型：文字列

(a-4) ページ名Ａの列４

・ページ名：Shop－Ａ

・対応列：列４

・列名：価格

・データ型：数値
(b-1) ページ名Bの列1
・ページ名：Shop－B
・対応列：列1
・列名：ショップ名
・データ型：文字列
(b-2) ページ名Bの列2
・ページ名：Shop－B
・対応列：列2
・列名：メーカ名
・データ型：文字列
(b-3) ページ名Bの列3
・ページ名：Shop－B
・対応列：列3
・列名：商品名
・データ型：文字列
(b-4) ページ名Bの列4
・ページ名：Shop－B
・対応列：列4
・列名：価格
・データ型：数値
次に、HTML文書に内在する情報の構成要素のローカルドメインを、図12に示すようにドメインテーブル154に定義する（ステップS130）。ショップAとショップBのショップ名、メーカ名、商品名については、各々任意の文字列であるため、特にローカルドメインを設定しない。一方価格については、図11の数値のデータ型の設定値を鑑み、ショップA、ショップBのローカルドメインを以下のように定義する。同時にこのローカルドメインをHTML文書項目テーブル153に登録する。
【0120】・ドメイングループ：価格
・ショップAのローカルドメイン：「￥」記号つき表現形式
・ショップBのローカルドメイン：数値と ”，” からなる表現形式
次に、ユーザ毎にユーザ入力ドメインとユーザ出力ドメインを、図13に示すようにユーザードメインテーブル155に以下のように定義する（ステップS140）。ユーザAは、ショップ名、メーカ名、商品名をHTML文書の表現形式で入力してそのままの表現形式での出力を要求するため、ユーザ入力ドメインとユーザ出力ドメインは設定しない。また、ユーザAは、価格ドメイングループについて、
・入力：「円」記号つき表現形式
・出力：「円」記号つき表現形式
を用いるものとする。このドメインの登録をドメインテーブル154にし、ユーザドメインの登録をユーザドメインテーブル155にする。尚、ユーザドメインは、ユーザ入力ドメインとユーザ出力ドメインで異なっていてもよい。
【0121】次に、ドメイン間の変換関数を、図14に示すようにドメイン変換関数テーブル156に定義する（ステップS150）。ドメインとして、数値と ”，” からなる表現形式、「円」記号つき表現形式、「￥」記号つき表現形式の3種類が存在するため、ユーザ入力ドメイン－ローカルドメイン、ユーザ出力ドメイン－ローカルドメインの相互変換用に以下の関数を以下のように作成し、ドメイン変換関数テーブル156に設定する。各変換関数は変換関数ライブラリ133に格納される。
【0122】(a) 数値と ”，” からなる表現形式から「円」記号つき表現形式への変換
・変換関数名：Num2Yen( )
・変換元ドメイン：数値と ”，” からなる表現形式
・変換先ドメイン：「円」記号つき表現形式
(b) 「円」記号つき表現形式から数値と ”，” からなる表現形式への変換
・変換関数名：Yen2Num( )
・変換元ドメイン：「円」記号つき表現形式
・変換先ドメイン：数値と ”，” からなる表現形式
(c) 数値と ”，” からなる表現形式から「￥」記号つき表現形式への変換
・変換関数名：Num2￥( )
・変換元ドメイン：数値と ”，” からなる表現形式
・変換先ドメイン：「￥」記号つき表現形式
(d) 「￥」記号つき表現形式から数値と ”，” からなる表現形式への変換
・変換関数名：￥2Num( )
・変換元ドメイン：「￥」記号つき表現形式
・変換先ドメイン：数値と ”，” からなる表現形式
(e) 「円」記号つき表現形式から「￥」記号つき表現形式への変換
・変換関数名：Yen2￥（）
・変換元ドメイン：「円」記号つき表現形式
・変換先ドメイン：「￥」記号つき表現形式
(f) 「￥」記号つき表現形式から「円」記号つき表現形式への変換
・変換関数名：￥2Yen（）
・変換元ドメイン：「￥」記号つき表現形式
・変換先ドメイン：「円」記号つき表現形式
（2）検索フェーズ
ユーザ「ユーザA」から以下の検索項目とその条件からなる簡易な構文の問い合わせ文が発行された場合の処理を例に説明する。
【0123】検索項目：ショップ名、メーカ名、商品名、価格
検索条件：価格 < 200,000円
まず、構文解析部12はユーザからの問い合わせを解析し、問い合わせ項目探索部131は指定された項目を検索（図6のステップS200）。指定された項目は「シ

ョップ名」、「メーカ名」、「商品名」、「価格」である。各項目と一致する列名を、ＨＴＭＬ文書項目テーブル１５３から探索すると、以下のレコードが得られる。
【０１２４】(a) ショップ名
・ページ名「Shop－Ａ」の対応列「列１」でデータ型「文字列」
・ページ名「Shop－Ｂ」の対応列「列１」でデータ型「文字列」
(b) メーカ名
・ページ名「Shop－Ａ」の対応列「列２」でデータ型「文字列」
・ページ名「Shop－Ｂ」の対応列「列２」でデータ型「文字列」
(c) 商品名
・ページ名「Shop－Ａ」の対応列「列３」でデータ型「文字列」
・ページ名「Shop－Ｂ」の対応列「列３」でデータ型「文字列」
(d) 価格
・ページ名「Shop－Ａ」の対応列「列４」でデータ型「数値」
・ページ名「Shop－Ｂ」の対応列「列４」でデータ型「数値」
次に、問い合わせ項目探索部１３１はすべての検索項目に対応する項目を保持するＨＴＭＬ文書名を探索する（ステップＳ２１０）。上記で得られた結果に対し、すべての検索項目に対応する項目を保持するＨＴＭＬ文書を探索すると、以下の２組が生成される。また、各組み合わせのＵＲＬをＨＴＭＬ文書テーブル１５１から取得する。
【０１２５】(A) 組み合わせ１
(a) 対象ページ名：Shop－Ａ
(b) 構成要素
・ショップ名：対応列「列１」でデータ型「文字列」
・メーカ名：対応列「列２」でデータ型「文字列」
・商品名：対応列「列３」でデータ型「文字列」
・価格：対応列「列４」でデータ型「数値」
(c) ＵＲＬ
http://www.shop－a.co.jp/products.html
(B) 組み合わせ２
(a) 対象ページ名：Shop－Ｂ
(b) 構成要素
・ショップ名：対応列「列１」でデータ型「文字列」
・メーカ名：対応列「列２」でデータ型「文字列」
・商品名：対応列「列３」でデータ型「文字列」
・価格：対応列「列４」でデータ型「数値」
(c) ＵＲＬ
http://www.shop－b.co.jp/shouhin.html
次に、問い合わせ変換部１３２は探索した項目に対応するユーザドメインとローカルドメインを取得する（ステップＳ２２０）。この探索した項目に対応するローカルドメインはＨＴＭＬ文書項目テーブル１５３を探索して得られる。ローカルドメインがある項目については、当該ローカルドメインのドメイングループをドメインテーブル１５４から探索し、当該ドメイングループに対するユーザドメインをユーザドメインテーブル１５５から取得する。結果として、以下の組み合わせを得る。
【０１２６】(A) 組み合わせ１
(a) 対象ページ名：Shop－Ａ
(b) 構成要素
・ショップ名：ローカルドメインなし
・メーカ名：ローカルドメインなし
・商品名：ローカルドメインなし
・価格：ローカルドメインは「￥」記号つき表現形式
ユーザ入力ドメインは「円」記号つき表現形式
ユーザ出力ドメインは「円」記号つき表現形式
(B) 組み合わせ２
(a) 対象ページ名：Shop－Ｂ
(b) 構成要素
・ショップ名：ローカルドメインなし
・メーカ名：ローカルドメインなし
・商品名：ローカルドメインなし
・価格：ローカルドメインは数値と”，”からなる表現形式
ユーザ入力ドメインは「円」記号つき表現形式
ユーザ出力ドメインは「円」記号つき表現形式
次に、問い合わせ変換部１３２はユーザ入力ドメインとローカルドメインが異なる項目に対し、ドメイン変換関数テーブル１５６から、変換元ドメインと変換先ドメインの一致する変換関数名を取得し、各ＨＴＭＬ文書のローカルドメインに変換する（ステップ２３０）。双方の組み合わせにおいて、価格の表現形式が、ローカルドメインとユーザ入力ドメインとで異なるので、変換元ドメインと変換先ドメインをキーに変換関数名をドメイン変換関数テーブル１５６から探索する。
【０１２７】(A) 組み合わせ１
変換元ドメイン：「円」記号つき表現形式
変換先ドメイン：「￥」記号つき表現形式
変換関数名：Yen2￥( )
(B) 組み合わせ２
変換元ドメイン：「円」記号つき表現形式
変換先ドメイン：数値と”，”からなる表現形式
変換関数名：Yen2Num( )
各々の組み合わせに対して変換関数を実行し、以下を得る。
【０１２８】(A) 組み合わせ１
Yen2￥(200,000円) ＝￥200,000
(B) 組み合わせ２
Yen2Num(200,000 円) ＝200,000
次に、問い合わせ変換部１３２は各ＨＴＭＬ文書アクセ

ス部１４に対する以下の検索文を生成する。
【０１２９】(A) 組み合わせ１
(a) 対象ページ名：Shop－Ａ
(b) 検索要求
検索項目：ショップ名、メーカ名、商品名、価格
検索条件：価格 <¥200,000
(B) 組み合わせ２
(a) 対象ページ名：Shop－Ｂ
(b) 検索要求
検索項目：ショップ名、メーカ名、商品名、価格
検索条件：価格 < 200,000
ＨＴＭＬ文書アクセス部１４はこれらの問い合わせ文により各ＨＴＭＬ文書毎に内在する情報の検索を実行し、ＨＴＭＬ文書を取得して検索結果を生成する（ステップＳ２４０）。ＨＴＭＬ文書処理部１３４はＵＲＬのリンク先から、各ＨＴＭＬ文書に内在する情報を、ＨＴＭＬ文書－表マッピングテーブル１５２に設定された情報に基づいて取り出し、検索条件が指定されていればフィルタリングを行い、以下の検索結果を得る。
【０１３０】(A) 組み合わせ１
(a) 対象ページ名：Shop－Ａ
(b) 検索結果
・ショップ名：ショップA、メーカ名：Maker Ａ、商品名：ＰＣ１、価格：¥170,000
・ショップ名：ショップA、メーカ名：Maker Ｂ、商品名：ＰＣ１０１、価格：¥198,000
(B) 組み合わせ２
(a) 対象ページ名：Shop－Ｂ
(b) 検索結果
・ショップ名：ショップB、メーカ名：Maker Ａ、商品名：ＰＣ１、価格：168,000
検索結果変換部１３５は、ユーザ出力ドメインとローカルドメインとが異なる項目がある場合、ドメイン変換関数を取得し、当該項目をユーザ出力ドメインに変換する（ステップＳ２５０）。上記の双方の組み合わせでは、価格が、ローカルドメインとユーザ出力ドメインとで異なるので、変換元ドメインと変換先ドメインをキーに変換関数名をドメイン変換関数テーブル１５６から探索する。
【０１３１】(A) 組み合わせ１
変換元ドメイン：「¥」記号つき表現形式
変換先ドメイン：「円」記号つき表現形式
変換関数名：¥2Yen( )
(B) 組み合わせ２
変換元ドメイン：数値と ", " からなる表現形式
変換先ドメイン：「円」記号つき表現形式
変換関数名：Num2Yen( )
各々の組み合わせに対して変換関数を実行し、以下の結果を得る。
【０１３２】(A) 組み合わせ１
¥2Yen（¥170,000)＝ 170,000円
¥2Yen（¥198,000)＝ 198,000円
(B) 組み合わせ２
Num2Yen(168,000)＝ 168,000円
最後に、ユーザーインターフェース部１１は以下の検索結果をユーザーに出力する（ステップＳ２６０）。
【０１３３】・ショップ名：ショップA、メーカ名：Maker Ａ、商品名：ＰＣ１、価格：170,000円
・ショップ名：ショップA、メーカ名：Maker Ｂ、商品名：ＰＣ１０１、価格： 198,000円
・ショップ名：ショップB、メーカ名：Maker Ａ、商品名：ＰＣ１、価格：168,000円
以上説明したように、第１の実施形態は、オープンなネットワーク上の複数ＨＴＭＬ文書に対し、各ＨＴＭＬ文書に内在する情報に関する各種の情報をメタデータとして管理する。このため、複数のＨＴＭＬ文書に内在する情報に対する一括の検索が実現でき、ＨＴＭＬ文書間の異種性による相違を解消した検索結果を生成することができる。同時に、各ＨＴＭＬ文書に内在する情報に関する情報をＨＴＭＬ文書ごと個別に管理するので、ＨＴＭＬ文書情報統合検索装置が検索対象とするＨＴＭＬ文書の追加、修正、削除の作業は当該ＨＴＭＬ文書だけに関して行えば足りる。このため、等比級数的に増加するＨＴＭＬ文書の本装置への検索対象としての取り込みが容易となる。
【０１３４】また、各ＨＴＭＬ文書からの検索結果は、項目ごとに条件処理可能な項目データとして得られるので、ＨＴＭＬ文書処理部１３４は、各ＨＴＭＬ文書の複数の検索結果を適宜マージして１つの検索結果とし、この１つの検索結果を必要に応じて条件処理することができる。
【０１３５】このように、第１の実施形態によれば、オープンなネットワークに散在する複数のＨＴＭＬ文書に対して該複数のＨＴＭＬ文書に内在する情報の文書構造、構成要素、表現形式等が互いに異なってもこれら複数の文書を跨った情報検索を実現し、ＨＴＭＬ記述上の差異をユーザ毎の統一形式に変換して一括して検索結果を返却することができる。従って従来に比較して、人手による多くの時間や労力が不要となり、検索効率が画期的に向上する。第１の実施形態は、例えば「ある製品を最安値で販売している店の名前と価格を求める」というようなエレクトロニック・コマースにおける柔軟な商品情報検索に利用可能であり、公正なエレクトロニック・コマースの市場の活性化に貢献し得る。
【０１３６】第２の実施形態図１５から図３６を参照して、本発明に係る半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびにそのプログラムを格納する記録媒体の第２の実施形態であるインターネット情報統合検索装置を詳細に説明する。
【０１３７】第２の実施形態は、オープンなネットワー

クに複数の情報検索装置（サーチエンジン）が散在する環境で、固有の入力フォームを持つ複数のサーチエンジンに対して各サーチエンジンの文書構造、入力フォームの必須入力項目、表現形式が互いに異なっていても、サーチエンジンを跨って条件指定を含む情報検索を行い、これら入力フォームの差異を解消して全サーチエンジンから検索結果を一括して取得することを実現するものである。

【０１３８】尚、第２の実施形態で用いられる表現形式の概念およびこれに関連する用語は、第１の実施形態と同様である。例えば、エリア名に対しても、「神奈川県」、「神奈川」と、ＨＴＭＬ文書の記述者や検索を実行するユーザによって、様々な表現形式が存在する。

【０１３９】例えば、エリアについて、神奈川県は「県」つき表現形式で１つのドメインであり、ジャンルについて、中華料理は「料理」つき表現形式で１つのドメインである。ドメイングループとしては、エリア、ジャンル、等がある。あるユーザが「神奈川県」、「中華料理」と入力する場合、ユーザ入力ドメインは「県」つき表現形式であり、ジャンルは「料理」つき表現形式である。あるユーザの出力が「神奈川県」、「中華料理」である場合、ユーザ出力ドメインは「県」つき表現形式であり、ジャンルは「料理」つき表現形式である。ＨＴＭＬ文書から抽出した検索結果が「神奈川県」である場合、ローカルドメインは「県」つき表現形式である。

【０１４０】尚、同一ドメイングループ内でユーザ入力ドメイン、ユーザ出力ドメイン、ローカルドメインが相互に異なる場合、第２の実施形態でも第１の実施形態同様、ドメイン変換関数を用いて、ドメイン間の差異を解消する。

【０１４１】図１５は、第２の実施形態に係るインターネット情報統合検索装置の構成を示すブロック図である。第２の実施形態は、図３の問い合わせ処理部１３を、さらに、入力必須項目探索部１３６と、検索パターン判定部１３７と、検索結果処理部１３８を具備する統合検索処理部１３０に置き換えた点において第１の実施形態の修正である。第２の実施形態に係るインターネット情報情報統合検索装置１０は、ユーザーインターフェース部１１と、構文解析部１２と、統合検索処理部１３０と、ＨＴＭＬ文書メタデータ記憶部１５と、ＨＴＭＬ文書メタデータ管理部１６と、ＨＴＭＬ文書アクセス部１４とから構成される。第２の実施形態に係る統合検索処理部１３０は、問い合わせ項目探索部１３１と、問い合わせ変換部１３２と、変換関数ライブラリ１３３と、入力必須項目探索部１３６と、検索パターン判定部１３７と、検索結果処理部１３８と、検索結果変換部１３５とを具備する。

【０１４２】尚、図３と同一の符号を付した箇所は、特に断らない限り第１の実施形態と同様であり、これらの説明は省略する。図１５において、ユーザーインターフェース部１１は、ユーザのアプリケーションプログラム３から入力される検索項目と検索条件からなる問い合わせ文を受け付ける。構文解析部１２は、ユーザインタフェース部１１で受け付けた問い合わせ文の構文を解析する。統合検索処理部１３０は、各サーチエンジンにより管理されるＨＴＭＬ文書に内在する項目を一括して検索する。統合検索処理部１３０中で、問い合わせ項目探索部１３１は、問い合わせ文中で指定された項目の所在を探索する。入力必須項目探索部１３６は、各サーチエンジンの入力フォーム上のデータ項目の不足をチェックして問い合わせ先のサーチエンジンを決定する。検索パターン判定部１３７は、問い合わせ文に応じた最適な検索パターンを判定して、この判定結果に従い問い合わせ文を最適化する。問い合わせ変換部１３２は、問い合わせ文のユーザ入力ドメインをローカルドメインに変換し、ＨＴＭＬ文書アクセス部１４が送出する問い合わせ文を生成する。検索結果処理部１３８は、ＨＴＭＬ文書アクセス部１４が受信したＨＴＭＬ文書から取得した情報に対し、問い合わせ文に応じた処理（検索項目の選択、検索条件によるフィルタリング等）を行う。検索結果処理部１３８はまた、抽出された情報に対して検索条件によるフィルタリング処理を行うとともに、上記で決定された検索パターンに応じてサーチエンジン側で行われた条件処理を抑止する。検索結果変換部１３５は、検索されたデータの表現形式をローカルドメインからユーザ出力ドメインに変換する。ＨＴＭＬ文書アクセス部１４は、生成された検索文を各サーチエンジンに送信し、オープンなネットワークに散在するＨＴＭＬ文書をサーチエンジンを介して取得する。このＨＴＭＬ文書に内在する情報が第２の実施形態により例えば表形式などの統一フォーマットに変換される。このＨＴＭＬ文書アクセス部１４には、通信網１９０を介してエンジン２３およびデータベース２４からなる複数のサーチエンジン２０－１，２０－２・・・が接続されている。ＨＴＭＬ文書メタデータ記憶部１５０は、各サーチエンジンの所在、各サーチエンジンの有するＨＴＭＬ文書の文書構造、ＨＴＭＬ文書に内在する情報の表現形式や構成要素などの各サーチエンジンに関する情報を記憶し管理する。ＨＴＭＬ文書メタデータ管理部１６は、ＨＴＭＬ文書メタデータ記憶部１５０に対する各種情報の入力／削除／変更を行う。システム管理者は、例えばエディタなどに実装されるＨＴＭＬ文書メタデータ管理部１６を介して、ＨＴＭＬ文書メタデータを登録・管理する。

【０１４３】図１６は、第２の実施形態に係るＨＴＭＬ文書メタデータ記憶部１５０が保有するテーブルの詳細を示す。図４に示す第１の実施形態のＨＴＭＬ文書メタデータ記憶部１５が有する各ＨＴＭＬ文書の所在情報を管理するＨＴＭＬ文書テーブル１５１と、各ＨＴＭＬ文書に内在する構成要素を表形式に変換するための情報を管理するＨＴＭＬ文書－表マッピングテーブル１５２

と、各項目ごとにこの項目の属性を管理するＨＴＭＬ文書項目テーブル１５３と、各ドメインの表現形式を管理するドメインテーブル１５４と、ユーザーごとに入力ドメインおよび出力ドメインを管理するユーザドメインテーブル１５５と、ドメイン変換関数を管理するドメイン変換関数テーブル１５６に加え、さらに各サーチエンジンの入力フォーム中の入力必須項目を管理する入力必須項目テーブル１５７とにより第２の実施形態のＨＴＭＬ文書メタデータ記憶部１５０は構成される。また検索パターン判定部１３７は、図２８に示すような内部に各サーチエンジンへの検索パターンを決定して検索文を各サーチエンジンごとに最適な問い合わせ文に変換するための検索パターンマトリックステーブルを具備する。あるいはこの検索パターンマトリックステーブルは、ＨＴＭＬ文書メタデータ１５０に含まれて構成されてもよい。

【０１４４】次に、第２の実施形態に係るインターネット情報統合検索装置１０の処理手順の詳細および各テーブルへの設定例を説明する。第２の実施形態の処理手順は、図１９に示す検索を実行する前に表現形式等の準備を行う準備フェーズと、図２９に示す検索を実行する検索フェーズの２段階のフェーズがある。

【０１４５】図１７（Ａ）、図１７（Ｂ）、図１７（Ｃ）に示すサーチエンジンの入力フォームが存在する場合の例で各フェーズを説明する。図１８には、図１７（Ｂ）のＰｅｇｅ－Ｂの入力フォームに対応するＨＴＭＬ記述を示す。

【０１４６】（１）準備フェーズ

図１９に示す準備フェーズではまず、ＨＴＭＬ文書項目テーブル１５３を例えば図２０に示すように設定する（ステップＳ３００）。ＨＴＭＬ文書項目テーブル１５３は、各サーチエンジン入力フォームの項目について、以下の項目を管理する。図２０で、ページ名は各サーチエンジンの入力フォーム名を示す。対応列は、ＨＴＭＬ文書－表マッピングテーブル１５２との対応付けを行う。データ項目名は、サーチエンジン入力フォームに内在する項目を示す。「項目指定可能」とは、当該項目がこのサーチエンジンの検索結果から取得できるか否かを示す。「条件指定可能」とは、当該項目がこのサーチエンジンによる検索の際に条件指定可能か否かを示す。データ型は、数値型、文字列型等のデータの処理タイプを示す。このデータ型はフィルタリング処理時のデータの評価方法として使用する。Nameタグは、選択形式となっている項目が有するNameタグを示す。ローカルドメインは、当該列が属するドメインを示す。

【０１４７】次に、ＨＴＭＬ文書テーブル１５１を、例えば図２１に示すように設定する（ステップＳ３１０）。ＨＴＭＬ文書テーブル１５１は、各サーチエンジン入力フォームの所在情報として、以下の項目を管理する。図２１で、ページ名は各サーチエンジンの入力フォーム名を示す。サーチエンジンＵＲＬは、各サーチエンジンの所在情報となるＵＲＬを示す。

【０１４８】次に、ＨＴＭＬ文書－表マッピングテーブル１５２を、例えば図２２に示すように設定する（ステップＳ３２０）。ＨＴＭＬ文書－表マッピングテーブル１５２は、各サーチエンジンから返却されるＨＴＭＬ文書に内在する情報の表形式への対応情報として、以下の項目を管理する。図２２で、ページ名は、各サーチエンジンの入力フォーム名を示す。「レコード開始」とは、取得されたＨＴＭＬ文書中での結果内容が開始される行をタグ情報を用いて示す。列１から列５は、取得されたＨＴＭＬ文書中の、検索結果とすべきデータ項目に対応する箇所をタグ情報を用いて定義する。列１から列５のそれぞれは、図２０のＨＴＭＬ文書項目テーブル１５３のページ名Ｐａｇｅ＿Ａの対応列「列１」から「列５」と対応する。次に、ドメインテーブル１５４を、例えば図２３に示すように設定する（ステップＳ３３０）。ドメインテーブル１５４は、ＨＴＭＬ文書項目テーブル１５３でローカルドメインを設定した列について、このローカルドメイン情報として、同じ意味を表すドメインの集まりであるドメイングループと、１つの表現の集まりであるドメインを管理する。

【０１４９】次に、ドメイン変換関数テーブル１５６を、例えば図２４に示すように設定する（ステップＳ３４０）。ドメイン変換関数テーブル１５６は、ドメイン変換関数情報として、以下の項目を管理する。図２４で変換関数名は、特定のドメインから特定のドメインへ変換するための関数の名前を示す。ドメイングループは同じ意味を表すドメインの集まりを示す。変換元ドメインはドメイン関数に対し、入力するドメインを示す。変換先ドメインはドメイン関数から出力されるドメインを示す。ライブラリ名はドメイン変換を実現する変換関数ライブラリ１３３のファイル名を示す。

【０１５０】次に、ユーザドメインテーブル１５５を、例えば図２５に示すように設定する（ステップＳ３５０）。ユーザードメインテーブル１５５は、ユーザがドメイングループ毎に、どのような入力ドメイン、出力ドメインを指定するかを以下の項目により管理する。図２５でユーザ名は、検索要求を行うユーザの名前を示す。ユーザ入力ドメインは、ユーザがあるドメイングループに対しどのようなドメインで入力するのかを示す。ユーザ出力ドメインは、ユーザがあるドメイングループから、どのようなドメインで出力されるかを示す。

【０１５１】次に、入力必須項目テーブル１５７を、例えば図２６に示すように設定する（ステップＳ３６０）。サーチエンジンによっては、入力フォーム中で入力を必須とされている項目がある。入力必須項目テーブル１５７は、この入力必須項目を、以下の項目により管理する。図２６でページ名は、各サーチエンジンの入力フォーム名を示す。入力必須項目は、サーチエンジンに対し、必ず入力する必要のある項目名を示す。

【0152】(2)検索フェーズ
図29は、第2の実施形態の検索検索実行時のフローチャートを示す。
【0153】ユーザが例えば「神奈川県にある和食料理の店」の「店名」と「電話番号」について調べたい場合の第2の実施形態の検索処理を、以下のSQLのSELECT文とWHERE 文のみからなる簡易な構文の問い合わせ文が入力された場合の例で説明する。

SELECT 店名、電話番号 WHERE エリア=”横浜市”and ジャンル=”和食料理” (1-1)

問い合わせ項目探索部131は、図20のHTML文書項目テーブルを参照し、検索項目および検索条件項目をデータ項目名に含むサーチエンジンを探索することにより、データ項目の所在を探索する(ステップS410)。図30にこのサーチエンジン探索結果を示す。
【0156】次に、問い合わせ項目探索部131は、ステップS410の結果からHTML文書テーブル151を参照し、「店名」、「電話番号」、「エリア」、「ジャンル」のすべての項目を満たすページを特定する(ステップS420)。この時点ではPage-A, Page-B, Page-Cが検索候補サーチエンジンとなる。
【0157】入力必須項目探索部136は、入力必須項目テーブル157を参照し、各サーチエンジンの必須項目をチェックして検索候補サーチエンジンを絞り込む

SELECT 店名、電話番号 WHERE エリア=”横浜市” (1-2)

の問い合わせ文が入力された場合には、問い合わせ項目探索部131においては、HTML文書項目テーブル153を参照することにより、Page-A、Page-B、Page-Cはいずれも項目「店名」、「電話番号」、「エリア」を含むため、検索候補サーチエンジンとされる。
【0159】次に、入力必須項目探索部136では、以下のように検索候補サーチエンジンが絞り込まれる。Page-Aは「ジャンル」を入力必須項目とする。これは、Page-Aに対する検索では「ジャンル」という項目の指定が必須であって、指定されない場合には検索できないことを意味する。問い合わせ条件(where句)には、「ジャンル」が指定されていないため、Page-Aは入力必須項目探索部136において検索対象から除外される。
【0160】Page-Cに対する検索では、「エリア」と「ジャンル」の両方の指定が必須であるため、検索対象から除外される。
【0161】一方、Page-Bの入力必須項目である「エリア」は問い合わせ条件(where句)で指定されているため、Page-Bは検索対象として選択される。
【0162】他方、入力必須項目を持たないサーチエンジンに対して上記(1-2)の問い合わせを行う場合には、このサーチエンジン(ページ)は入力必須条件がないため、問い合わせ条件(where句)が指定されていても検索できる。従って、入力必須項目探索部136で検索対象サーチエンジンとして選択される。
【0163】この時点での問い合わせ文(1-1)に基づく各サーチエンジンへのSQL文はそれぞれ以下の通りである。
【0164】

Page-A:
SELECT 店名、電話番号 WHERE エリア=”横浜市”and ジャンル=”和食料理” (2-1)

Page-B:
SELECT 店名、電話番号 WHERE エリア=”横浜市”and ジャンル=”和食料理” (2-2)

Page-C:
SELECT 店名、電話番号 WHERE エリア=”横浜市”and ジャンル=”和食料理” (2-3)

次に、検索パターン判定部137は、図28の検索パターンマトリックスを参照して検索の処理方法を決定する

(ステップS440)。ここで、この検索パターンマトリックスを説明する。図27は第2の実施形態に係るインターネット情報統合検索装置と各サーチエンジンとの簡略化した関係を示す。ユーザーから入力される問い合わせ文の処理手順には、図27中の(a),(b),(c)の3つの検索パターンがある。(a)パターンは検索要求を未処理で返却する。(b)パターンは各サーチエンジンで条件処理を行う。(c)パターンは各サーチエンジンで条件処理を行ったのち、その結果を第2の実施形態に係る装置10でフィルタリング処理する。検索パターンマトリックスは、各検索文中の検索項目がそれぞれ上記3パターンのいずれに属するかを判定するために用いられる。検索パターン判定部137は、この図28に示す検索パターンマトリックスを用いて検索を実現するための戦略を決定する。図28で、検索要求の「項目」は検索すべき項目として例えばSQLのｓｅｌｅｃｔ句で指定された項目である。検索要求の「条件」は検索要求の検索条件として例えばSQLのｗｈｅｒｅ句で指定された項目である。エンジン(サーチエンジン)の「項目」は各サーチエンジンが検索結果として返す項目である。エンジンの「条件」は例えば各サーチエンジンの入力フォームにより規定される、各エンジンが検索要求として受け付ける条件である項目である。尚、エンジンの「項目」はHTML文書項目テーブル153の「項目指定可能」の欄の値を、エンジンの「条件」はHTML文書項目テーブル153の「条件指定可能」の欄の値を示す。処理パターン中の「検索条件値をそのまま返却」とは、指定された検索項目を処理することなく条件値を戻すことを示す。「情報源から返却されたものを返却」とは、指定された検索項目に対応してサーチエンジンから戻された結果を戻すことを示す。「サーチエンジンで処理」とは、指定された検索条件をサーチエンジンで処理することを示す。「装置でフィルタリング」とは、指定された検索条件に対してサーチエンジンから戻された検索結果を、検索結果処理部138で条件処理することを示す。

【0165】例えば、問い合わせ文(1-1)の場合、「店名」はselect句で指定されており、where句では指定されていない。この項目”店名”は図28の”検索要求”の「項目」欄が○で「条件」欄が×の行に相当する。一方、例えば図17(A)のサーチエンジンの入力フォームｐａｇｅ＿Ａは、図20のHTML文書項目テーブル153を参照すると、「店名」を条件として受け取り、かつ検索結果として返すことができる。このため図28のエンジンの「項目」、「条件」欄はともに○と定まる。従って、項目「店名」は図28の上から4行目のレコードに対応する。従って「店名」のＰａｇｅ＿Ａに対する処理パターンは、エンジンから返されたデータを項目として返し、SQLで条件を指定していないため条件は処理しないことがわかる。

【0166】一方、「エリア」はselect句で指定されておらず、where句で指定されている。この項目「エリア」は図28の「検索要求」の「項目」欄が×で「条件」欄が○の行に相当する。一方、例えば図17(A)のＰａｇｅ＿Ａは、図20のHTML文書項目テーブル153を参照すると、「エリア」を条件として受け取ることはできないが、「エリア」を検索結果として返すことができる。このため図28のエンジンの「項目」欄は○、「条件」欄は×と定まる。従って、項目「エリア」は図28の上から8行目の行にレコードに対応する。従って「エリア」のＰａｇｅ＿Ａに対する処理パターンは、SQLでselect句に指定がないため項目としては返さず、エンジンでは条件として処理できないため検索結果処理部138でフィルタリング処理して返すことがわかる。(1-1)の問い合わせ文で指定されている他の項目「電話番号」、「ジャンル」についてもＰａｇｅ＿Ａを対象として上記の当てはめ処理を行うことで、図28から図31のマトリックスが導出される。

【0167】図31は、検索要求とPage-Aに指定可能な項目および条件項目を各データ項目毎に判定された処理内容を示す。図31から、「サーチエンジンで処理」の欄に基づき、「ジャンル」を検索条件としてPage-Aに送信すべきことがわかる。また「装置でフィルタリング」の欄に基づき、Page-Aからの検索結果を「エリア」の条件でフィルタリング処理すべきことがわかる。また「情報源から返却されたものを返却」の欄に基づき、「店名」、「電話番号」はPage-Aからの送信結果をそのまま返却すべきことがわかる。

【0168】Ｐａｇｅ＿Ａに対して、問い合わせ文(1-1)により検索する場合、Ｐａｇｅ＿Ａでは「店名」と「ジャンル」が条件として指定可能だが、問い合わせ文(1-1)では「ジャンル」のみ条件指定されている。このため、「ジャンル」を”和食料理”としてＰａｇｅ＿Ａのサーチエンジンには問い合わせ、かつ検索結果処理部138でのフィルタリング処理により、「エリア」が”横浜市”である「店名」、「電話番号」のデータを選択して検索結果とする。従って、Page-Aへの検索は上記のパターンCであり、問い合わせ文(2-1)は以下のように変換される。

【0169】

フィルタリング条件:「エリア」=”横浜市”

SELECT 店名、電話番号 WHERE ジャンル=”和食料理” (3-1)

同様の手順で、Page-B,Page-Cに対する各問い合わせ文が生成される。図32は、Page-Bについて判定された処理内容を示す。図32から、「情報検索装置で処理」の欄に基づき、「エリア」を検索条件としてPage-Bに送信すべきことがわかる。「装置でフィルタリング」の欄に基づき、Page-Bからの検索結果を「ジャン

ル」の条件でフィルタリング処理すべきことがわかる。「情報源から返却されたものを返却」の欄に基づき、「店名」、「電話番号」はPage－Bからの送信結果をそのまま返却すべきことがわかる。従って、Page－Bへの

フィルタリング条件：「ジャンル」＝”和食料理”

SELECT　店名、電話番号　WHERE エリア＝”横浜市”　（３－２）

図３３は、Page－Cについて判定された処理内容を示す。図３３から、「情報検索装置で処理」の欄に基づき、「エリア」、「ジャンル」を検索条件としてPage－Cに送信すべきことがわかる。「装置でフィルタリング」の欄に基づき、Page－Cからの検索結果にフィルタリング処理は行わないことがわかる。「情報源から返却

フィルタリング条件：なし

SELECT　店名、電話番号　WHERE エリア＝”横浜市”and ジャンル＝”和食料理”　（３－３）

次に、問い合わせ変換部１３２は、検索パターン判定部１３７が出力する各サーチエンジンへの問い合わせ文を、各サーチエンジンのローカルドメインに適合する検索文に変換する（ステップＳ４５０）。問い合わせ変換部１３２は、検索条件で指定された項目に対応するサーチエンジンの項目のうち、ローカルドメインが設定されている項目のユーザ入力ドメインとローカルドメインを、ＨＴＭＬ文書項目テーブル１５３およびユーザドメインテーブル１５５から図３４に示すように取得する。ユーザ入力ドメインとローカルドメインが異なる項目について、ドメイン変換関数テーブル１５６から変換関数ライブラリ１３３中の関数情報を取得しこれらの項目をローカルドメインの表現形式に変換する。例えば、Page－Bのデータ項目名「エリア」の場合、ローカルドメインは”Page－Ｂ－City”である。このドメイングループに対するユーザー入力ドメインは、ユーザードメインテーブル１５５およびドメインテーブル１５４より、ドメインSHITSUKIである。このため、問い合わせ変換部１３２は、ドメイン変換関数テーブル１５６を参照し”Shi2ValueB（）”を用いて”横浜市”を入力フォーム中の選択リスト中の７番目の項目であることを示す”07”に変換する。

【０１７２】同様に、Page－Cのデータ項目名「ジャンル」の場合、ローカルドメインは”Page－Ｃ－Dishes”である。このドメイングループに対するユーザー入力ドメインは、ユーザードメインテーブル１５５およびドメインテーブル１５４より、ドメイン”RYOURITSUKI ”である。このため、問い合わせ文変換部１３２は、ドメイン変換関数テーブル１５６を参照し”Ryouri2ValueC（）”を用いて”和食料理”を選択リスト中の１番目の項目であることを示す”１”に変換する。

【０１７３】この時点で生成されている各サーチエンジンへの問い合わせ文および検索結果処理部１３８でのフィルタリング条件は、以下の通りである。

【０１７４】

Page－Ａ：

フィルタリング条件：「エリア」＝”横浜市”

SELECT　店名、電話番号　WHERE ジャンル＝”和食料理”（４－１＝３－１）

Page－Ｂ：

フィルタリング条件：「ジャンル」＝”和食料理”

SELECT　店名、電話番号　WHERE エリア＝”07”　（４－２）

【０１７５】

検索文（４－２）では、エリア＝”横浜”はエリア＝”07”に変換されている。

Page－Ｃ：

SELECT　店名、電話番号 FROM Page－Ｃ

WHERE　エリア＝”横浜市”and ジャンル＝”１”　（４－３）

検索文（４－３）では、ジャンル＝”和食料理”はジャンル＝”１”に変換されている。

【０１７６】次に、ＨＴＭＬ文書アクセス部１４は、ステップＳ４６０で得られた問い合わせ文に基づいて各サーチエンジン固有の以下の検索要求をそれぞれ発行する。各サーチエンジンではそれぞれ検索処理が実行される（ステップＳ４７０）。

【０１７７】

Page－Ａ：

フィルタリング条件：「エリア」＝”横浜市”

” GET http://www. Page−a.co.jp/search-shop.cgi?category＝和食料理 htt
p/1.0 ”　（5−1）

Page−B：フィルタリング条件：「ジャンル」＝”和食料理”

” GET http://www. Page−b.co.jp/search-shop.cgi?area＝07 http/1.0 ”
（5−2）

Page−C：

” GET http://www. Page−c.co.jp/search-shop.cgi?area＝横浜市 & categor
y ＝1 http/1.0”　（5−3）

次に、検索結果処理部１３８は、各サーチエンジンから返却された各ＨＴＭＬ文書に内在する情報を、ＨＴＭＬ文書−表マッピングテーブル１５２に設定された情報に基づいて抽出する（ステップＳ４７５）。図３５（Ａ）はPage−Ｂの検索結果のＨＴＭＬ文書のブラウザでの表示例を示す。図３５（Ｂ）は図３５（Ａ）の表示に対応するＨＴＭＬ記述を示す。以下に各サーチエンジンから得られた検索結果を示す。

【０１７８】

(a) 対象ページ名：Page−Ａ
フィルタリング条件：「エリア」＝”横浜市”
検索結果：
・店名：A1　エリア：横浜市
　　電話番号：(045) ＊＊＊−＊＊＊＊
・店名：A2　エリア：横須賀市
　　電話番号：(0468)＊＊−＊＊＊＊　（6−1）
(b) 対象ページ名：Page−Ｂ
フィルタリング条件：「ジャンル」＝”和食料理”
検索結果
・店名：B1　ジャンル：和食料理
　　電話番号：045−＊＊＊−＊＊＊＊
・店名：B2　ジャンル：中華料理
　　電話番号：045−＊＊＊−＊＊＊＊
・店名：B3　ジャンル：中華料理　（6−2）
　　電話番号：045−＊＊＊−＊＊＊＊
(c) 対象ページ名：Page−Ｃ
フィルタリング条件：なし
検索結果
・店名：C1　電話番号：045−＊＊＊−＊＊＊＊
・店名：C2　電話番号：045−＊＊＊−＊＊＊＊　（6−3）

次に、検索結果処理部１３８は、図２８の検索パターンマトリックスでフィルタリング処理を行うと規定されている項目について（ステップＳ４８０Ｙ）、各サーチエンジンからの検索結果をフィルタリング処理する（ステップＳ４９０）。ここで、Page−Aは「エリア」を評価せず、Page−Ｂは「ジャンル」を評価しない。従って、これらの結果から、「エリア」＝”横浜市”、「ジャンル」＝”和食料理”の条件を満たす結果を以下のように抽出する。

【０１７９】

(a) 対象ページ名：Page−Ａ
フィルタリング結果
・店名：A1　電話番号：(045)＊＊＊−＊＊＊＊　（7−1）
(b) 対象ページ名：Page−Ｂ
フィルタリング結果
・店名：B1　電話番号：045−＊＊＊−＊＊＊＊　（7−2）
(c) 対象ページ名：Page−Ｃ
フィルタリング結果
・店名：C1　電話番号：045−＊＊＊−＊＊＊＊
・店名：C2　電話番号：045−＊＊＊−＊＊＊＊　（7−3＝6−3）

次に、検索結果変換部１３５は、検索項目で指定された項目で、ローカルドメインが指定されている項目のユーザ出力ドメインとローカルドメインをＨＴＭＬ文書項目テーブル１５３、ドメインテーブル１５４およびユーザドメインテーブル１５５から図３６に示すように取得する。検索結果変換部１３５は、ユーザ出力ドメインとローカルドメインが異なる項目に対し、ドメイン変換関数テーブル１５６から関数情報を取得しユーザー出力ドメインに変換する（Ｓ５００）。Page－Ａのデータ項目名「電話番号」の場合、ローカルドメインとユーザ出力ドメインが一致しているため、変換は行わない。一方、Page－Ｂ，Page－Ｃのデータ項目名「電話番号」の場合、ローカルドメインは Tel－Bar であるのに対し、出力ドメインは Tel－Paren である。このため、検索結果変換部１３５は、ドメイン変換関数テーブル１５６を参照してBar2Paren() を用い「 045－＊＊＊－＊＊＊＊」を「(045) ＊＊＊－＊＊＊＊」に変換する。Page－Ｂ，Page－Ｃのローカルドメインは、以下のようにユーザー出力ドメインに変換される。

【０１８０】入力　：「 045－＊＊＊－＊＊＊＊」（ドメイン： Tel－Bar ）
ドメイン変換関数：Bar2Paren()
出力　：「(045) ＊＊＊－＊＊＊＊」（ドメイン： Tel－Paren ）
以上の処理により、ユーザーインターフェース部１１は、統一検索結果を以下のように返却する。ユーザー側のアプリケーションプログラムでは、これらの統一検索結果を表形式などの統一フォームで表示する（ステップＳ５１０）。

【０１８１】
・店名：A1　電話番号：(045) ＊＊＊－＊＊＊＊
・店名：B1　電話番号：(045) ＊＊＊－＊＊＊＊
・店名：C1　電話番号：(045) ＊＊＊－＊＊＊＊
・店名：C2　電話番号：(045) ＊＊＊－＊＊＊＊

以上説明したように、第２の実施形態によれば、オープンなネットワークに散在する複数のサーチエンジンに対して検索を要求する場合、各サーチエンジン固有の入力フォームのオブジェクトを個別に管理することにより、異種の各サーチエンジンへのインターフェースの相違を解消して、複数の任意の入力項目に対応した柔軟な検索が可能となる。このため、サーチエンジンの異種性に起因する複数のサーチエンジンから返却されたＨＴＭＬ文書に内在する情報の文書構造、表現形式、入力フォームの差異を解消し、統一フォーマットによる検索結果の閲覧が可能となる。従って、検索効率が向上するとともに、ネットワークへの無効なトラフィックを軽減する。また、各サーチエンジンの入力フォームを個別に登録、管理するのでＨＴＭＬ文書メタデータの維持管理が容易に行える。

【０１８２】第３の実施形態図３７から図５０を参照して、本発明に係る半構造化文書情報統合検索装置および半構造化文書情報抽出装置、その方法、ならびにそのプログラムを格納する記録媒体の第３の実施形態であるＨＴＭＬ文書情報抽出装置を詳細に説明する。

【０１８３】第３の実施形態は、オープンなネットワークに散在するＨＴＭＬ文書に対し、各ＨＴＭＬ文書に内在する情報を、項目別に抽出する情報検索を実現する。第３の実施形態は、図３のＨＴＭＬ文書処理部１３４を、テンプレート解析部１３４１と、ＵＲＬ－テンプレート対応表１３４２と、テンプレート処理部１３４３とにより構成した点において、第１の実施形態の修正である。尚、第３の実施形態は、図３および図１５の構文解析部１２，問い合わせ項目探索部１３１，問い合わせ変換部１３２，ＨＴＭＬ文書メタデータ記憶部１５、１５０、ＨＴＭＬ文書メタデータ管理部１６等を備えることにより上記の第１または第２の実施形態と適宜組み合わせて実施されてもよく、また図３７に示す構成により単独で実施されてもよい。

【０１８４】第３の実施形態は、ＨＴＭＬ文書から情報を項目別に抽出するために必要となるＨＴＭＬ文書の所在情報と、ＨＴＭＬ文書に内在するデータの文書構造情報とを、各ＨＴＭＬ文書ごとに設定し、これらの情報を用いてＨＴＭＬに内在する情報を項目別に抽出する。具体的には所在情報は、各ＨＴＭＬ文書のＵＲＬとして個別に管理される。ＰＲＯＸＹの情報は、ＰＲＯＸＹ設定ファイル中のＰＲＯＸＹサーバ名、ＰＲＯＸＹポート番号として管理されてもよい。文書構造情報は、ＨＴＭＬ文書中の表、リスト、箇条書きなどの部分構造に関する情報であり、例えば抽出すべき項目がタグやスラッシュなどのデリミタで区切られているという情報である。この文書構造情報には、各項目の列、データ型などの各項目の各種属性情報を含む。この文書構造情報は、テンプレートファイル中の項目名、抽出テキスト規定部、項目名のデータ型等として管理される。このデータ型は例えば文字型、数値型などの値を持ち、各項目を処理するための付加情報として定義される。各テンプレートファイルと検索すべきＨＴＭＬ文書のＵＲＬまたはファイル名は、ＵＲＬまたはファイル名と、テンプレートファイル名とを有するＵＲＬ－テンプレート対応表により対応付けられる。また、各ＨＴＭＬ文書は、テンプレートファイル中の抽出テキスト規定部分が規定する表形式への対応情報を参照することにより、表形式などの統一フォーマットに変換される。尚、テンプレートファイルは、図４および図１６のＨＴＭＬ文書－表マッピングテーブル１５２およびＨＴＭＬ文書項目テーブル１５３に対応する。

【０１８５】第３の実施形態は、これらのＰＲＯＸＹ設定ファイル、ＵＲＬ－テンプレート対応表、テンプレートファイルを参照し、ユーザがＵＲＬまたはファイル名を指定すると、例えばＵＲＬ指定時にはＰＲＯＸＹ設定

ファイルを参照してＨＴＭＬ文書を取得した後、ＵＲＬ－テンプレート対応表を参照して該当するテンプレートファイル名を取得し、取得されたＨＴＭＬ文書を先頭から順番に１行または複数行単位でスキャンして、テンプレートファイルに記述される抽出テキスト規定部分と比較した結果に基づいて項目を抽出する。この項目抽出の際には、テンプレートファイル中で次ページへのリンクの有無を検証し、次ページへのリンクが存在する場合、このリングがなくなるまで次ページのＵＲＬもしくはファイル名を抽出してこのページの項目を抽出する処理を繰り返し行う。テンプレートファイルを参照して項目のマッチング判定を行うことにより、ＨＴＭＬ文書に内在する情報が表形式にマッピングされて項目単位に抽出される。第３の実施形態は、この抽出された各項目のデータをテンプレートファイルで規定されるデータ型に整形し、ユーザーに項目名と整形済み項目情報を返却する。従来の技術と比較して、ＨＴＭＬ文書中では本来文字型である文書の構成要素のデータ型を任意に規定することにより検索条件を用いて抽出された情報を条件処理できる。さらに、第１および第２の実施形態と同様、項目データの表現形式をユーザが所望する形式に加工できる。

【０１８６】図３７は、第３の実施形態に係るＨＴＭＬ文書情報抽出装置の構成を示すブロック図である。第３の実施形態に係るＨＴＭＬ文書情報抽出装置１００は、ユーザーアクセス部１１と、ＨＴＭＬ文書アクセス部１４と、ＰＲＯＸＹ設定ファイル１４１と、ＨＴＭＬ文書処理部１３４と、テンプレートファイル１３４５と、検索結果変換部１３５とで構成される。ＨＴＭＬ文書処理部１３４は、テンプレート解析部１３４１と、ＵＲＬ－テンプレート対応表１３４２と、テンプレート処理部１３４３とを有する。ＨＴＭＬ文書情報抽出装置１００は、ユーザーのアプリケーションプログラム３からの問い合わせ文３０１に基づき、ＰＲＯＸＹサーバー２を介してＨＴＭＬ文書にアクセスして、あるいは直接ローカルのＨＴＭＬ文書にアクセスして、これらＨＴＭＬ文書から得られた情報をテンプレート処理して検索結果３０２としてユーザーに返却する。

【０１８７】ＨＴＭＬ文書情報抽出装置１００は、複数のＨＴＭＬ文書がネットワーク上に散在する環境で、ＨＴＭＬ文書の所在、使用されるタグの種類、内包される構成要素の種類が異なっていても、ＨＴＭＬ文書から項目毎に情報を抽出するのに必要となる上記の各ＨＴＭＬ文書の所在情報、文書構造情報を各ＨＴＭＬ文書個別に設定することにより、ＨＴＭＬ文書からの所望する検索結果の表形式などの統一フォーマットでの抽出を実現するものである。

【０１８８】ＨＴＭＬ文書情報抽出装置１００のユーザアクセス部１１は、ユーザからの問い合わせ文をアプリケーションプログラム３から受信し、ＨＴＭＬ文書アクセス部１４に送信する。ＨＴＭＬ文書アクセス部１４は、ユーザアクセス部１１から受信したＵＲＬまたはファイル名に基づいて必要に応じＰＲＯＸＹ設定ファイル１４１を参照して、ＨＴＭＬ文書４－１、４－２を取得する。この取得されたＨＴＭＬ文書４－１、４－２をテンプレート解析部１３４１に送信する。ＨＴＭＬ文書アクセス部１４はまた、取得されたＨＴＭＬ文書がリンク情報を含む場合には、テンプレート解析部１３４１が抽出したリンク先ＵＲＬに基づいて、必要に応じＰＲＯＸＹ設定ファイル１４１を参照して、ＨＴＭＬ文書４－１、４－２を取得する。ＰＲＯＸＹ設定ファイル１４１は、図３９に示すように、ＨＴＭＬ文書４－１、４－２を取得するために必要なＰＲＯＸＹサーバの所在情報であるＰＲＯＸＹサーバ名、ＰＲＯＸＹポート番号を規定したファイルであり、ＨＴＭＬ文書アクセス部１４により参照される。テンプレートファイル１３４５は、図４０に示すように、ＨＴＭＬ文書４－１、４－２から項目として抽出可能な部位および抽出項目を抽出テキスト規定部分に規定し、各抽出項目のデータ型を規定するファイルであり、テンプレート解析部１３４１により参照される。ＵＲＬ－テンプレート対応表１３４２は、受信したＵＲＬ情報を元に、当該ＵＲＬまたはファイル名がどのテンプレートと対応しているかを管理するファイルであって、テンプレート解析部１３４１によって参照される。テンプレート解析部１３４１は、ＵＲＬ－テンプレート対応表１３４２を参照して、問い合わせ文に対応するテンプレートファイル１３４５の名称を取得する。同時に、このテンプレートファイル名を有するテンプレートファイル１３４５を参照し、取得されたＨＴＭＬ文書の抽出可能な部位、抽出すべき項目、抽出すべき項目のデータ型等を解析、取得し、テンプレート処理部１３４３へ送信する。この際テンプレートファイル１３４５上でリンク先ＵＲＬの有無も判断され、テンプレート解析部１３４１はリンク先が存在する場合にはＨＴＭＬ文書アクセス部１４にリンク先ＵＲＬを送信してリンク先ＨＴＭＬ文書を取得する。テンプレート処理部１３４３は、テンプレート解析部１３４１から受信した抽出可能な部位、抽出すべき項目、抽出すべき項目のデータ型に基づいてＨＴＭＬ文書４－１，４－２から各項目を抽出する。検索結果変換部１３５は、テンプレート処理部１３４３から抽出されたデータおよびそのデータ型を受信し、データ型に沿った抽出データの変換処理を行う。この変換後の抽出データを検索結果３０２としてユーザインターフェース部１１に送出する。

【０１８９】なお、このＨＴＭＬ文書情報抽出装置１００は、第１および第２の実施形態と同様、ＣＰＵ、メモリ、入出力装置、外部記憶装置等からなるコンピュータと、該コンピュータに読み取られた際、このコンピュータを前記各手段として機能させるためのプログラムを記憶した媒体とによって実現することもできる。

【０１９０】ＰＲＯＸＹサーバ2は、ＨＴＭＬ文書情報

抽出装置１００から指定されることが可能なＨＴＭＬ文書取得の仲介を行うサーバであり、ＵＲＬによって指定されたＨＴＭＬ文書４－１をＨＴＭＬ文書情報抽出装置１００に返却する。ＨＴＭＬ文書４－１，４－２は、オープンなネットワーク上に散在するホームページを記述したタグ付きテキストファイルである。アプリケーションプログラム３は、ユーザからのＵＲＬまたはファイル名と、少なくとも検索項目を含む問い合わせ文を受け付け、ＨＴＭＬ文書情報抽出装置１００からの受信結果をユーザに出力する。

【０１９１】次に、第３の実施形態に係るＨＴＭＬ文書情報抽出装置１００の処理手順を説明する。第３の実施形態の処理手順は、図３８に示す検索を実行する前に表現形式等の準備を行う準備フェーズと、図４１に示す検索を実行する検索フェーズの２段階のフェーズがある。尚、図３８の準備フェーズの手順は管理者が適当なエディタ等を用いて作成・設定するものであり、ＨＴＭＬ文書情報抽出装置１００自体を動作させて行うものではない。

【０１９２】（１）準備フェーズ

図３８に示す準備フェーズでは、まず図３９に示すようにＰＲＯＸＹサーバが必要な場合（ステップＳ６００Ｙ）、ＰＲＯＸＹサーバ名、ＰＲＯＸＹポート番号を定義してＰＲＯＸＹ設定ファイル１７１が作成される（ステップＳ６０５）。次に、テンプレートファイルが作成される（ステップＳ６１０）。このテンプレートファイルには、複数のテンプレートファイル間で一意となるファイル名が与えられ、図４０に例として示すように以下の情報が記述される。

【０１９３】(a)抽出項目

この抽出項目は、図４０の「Word」キーワードに対応する。

【０１９４】ＨＴＭＬ文書から抽出したい情報として、抽出すべき項目名、抽出すべき項目のデータ型、抽出すべき項目に付け加える固定値を記述する。図４０でこのデータ型は、”１”が文字型を示す。尚、このデータ型には、”３”が数値型、”４”が文字列を追加する型等と所望する条件処理に応じて設定することができる。図４０のテンプレートファイルには、リンク先アドレス（ＵＲＬの相対パス）等が「NextURL」で始まる部分に記述されている。これらの抽出項目のデータ型及び抽出項目に付け加える固定値は、ユーザーに検索結果を返却する際に必要な情報を追加もしくは削除するために必要となる。

【０１９５】(b)抽出テキスト規定部分

この抽出テキスト規定部分は、図４０の「HtmlTemplate」部分に対応する。

【０１９６】抽出対象となるＷｅｂページより、抽出したい情報を含むＨＴＭＬ文書の一レコード分をコピーする。そのうち、取得したい情報部分を「＄抽出項目名＄」に置き換え、残った記述のうちレコードに依存している省略可能な記述を、省略記号「..」に置き換える。

【０１９７】また、同一ＨＴＭＬ文書内に異なるテーブルとして取り扱うべき情報が混在する場合、同一テーブルの最後を特定する文字列を記入する。図４０では、第１、第２および第３の表の項目がそれぞれ定義されている。

【０１９８】さらにリンク先のＵＲＬが存在する場合、リンク先ＵＲＬを特定する文字列を記入する。

【０１９９】次に、ＵＲＬ－テンプレート対応表を作成する（ステップＳ６２０）。各ＵＲＬまたはファイルに対し、図４１に示すように対応するテンプレートファイル名を記述したファイルを作成する。

【０２００】（２）実行フェーズ

図４２は、第３の実施形態が取得したＨＴＭＬ文書から所望する項目を抽出する実行フェーズの処理手順を示すフローチャートである。

【０２０１】まず、ユーザーインターフェース部１１は、ユーザーがアプリケーションプログラム３に入力したＵＲＬまたはファイル名と、検索項目を含む検索文を受け付ける（ステップＳ７００）。ＨＴＭＬ文書アクセス部１４は入力がＵＲＬの場合、ＰＲＯＸＹ設定ファイル１４１があればそれを参照してＨＴＭＬ文書４－１を取得する。入力がファイル名の場合、ローカルのＨＴＭＬ文書が指定される。ユーザアクセス部１１０により受信されたＵＲＬまたはファイル名とＰＲＯＸＹ設定ファイル１４１の記述内容に従って、ＨＴＭＬ文書アクセス部１４はＰＲＯＸＹサーバ２を介するか、直接ＨＴＭＬ文書を取得する。また、ＨＴＭＬ文書アクセス部１４は返却結果であるＨＴＭＬ文書４－１を受信する（ステップＳ７１０）。

【０２０２】テンプレート解析部１３４１は、ＵＲＬと対応するテンプレートファイルの有無を判定する。ユーザインターフェース部１１を介し受信したＵＲＬまたはファイル名を参照し、このＵＲＬまたはファイル名に対応するテンプレートファイル名を図４１のＵＲＬ－テンプレート対応表１３４２から探索する（ステップＳ７２０）。対応するテンプレートファイルが存在しない場合（Ｓ７２０Ｎ）、ユーザインターフェース部１１に対しエラーメッセージを送信する。一方存在すれば（Ｓ７２０Ｙ）、テンプレート解析部１３４１は、取得されたＨＴＭＬ文書に対応するテンプレート名のテンプレートファイル１３４５に記述されている抽出ルールを解析し（ステップＳ７３０）、抽出に必要な情報をテンプレート処理部１３４３に送信する。

【０２０３】テンプレート処理部１３４３は、テンプレートファイル１３４５の抽出ルールを用いて、ＨＴＭＬ文書４－１から実際に項目を抽出して表形式のデータを得る（ステップＳ７４０）。テンプレート処理部１３４３は、ステップＳ７３０の抽出ルール解析によりリンク

先ＵＲＬの有無を判定する（ステップＳ７５０）。リンク先のＵＲＬが取得された場合（ステップＳ７５０Ｙ）、リンク先ＵＲＬをＨＴＭＬ文書アクセス部１４に送信する。ＨＴＭＬ文書アクセス部１４により取得されたリンク先ＨＴＭＬ文書に対してステップＳ７３０～Ｓ７５０の処理を行う。

【０２０４】検索結果変換部１３５は、抽出された項目の項目データを、図４０のテンプレートファイル１３４５を参照することで、以下のデータ変換処理を行って加工する。

【０２０５】ａ）．抽出した情報をそのまま表示すべきデータ型の項目データに、変換は実施しない。

【０２０６】ｂ）．固定値を代入すべきデータ型の項目データには、ＨＴＭＬ文書中に存在しないが、項目として返却したい項目について検索結果変換処理部１３５が有する固定値を返却する。

【０２０７】ｃ）．取得情報からカンマを削除すべきデータ型の項目データには、数値情報中からカンマを削除する。

【０２０８】ｄ）．取得情報に追加すべきデータ型の項目データには、ＵＲＬの相対パス等、抽出項目に対し付加すべき固定値が存在する場合、当該固定値を付加する。

【０２０９】上記の処理で得られるすべての検索結果は、ユーザインターフェース部１１を介してアプリケーションプログラム３に送信され、表示される。

【０２１０】図４３～図４６は第３の実施形態による項目情報抽出の具体例を示すもので、図４３はＨＴＭＬ文書のWebブラウザでの表示例、図４４は図４３の表示に対応するＨＴＭＬ記述例（但し、その一部分）である。図４５は、図４３、図４４のＨＴＭＬ文書からの項目抽出を行うためのテンプレートファイル１７１を示すもので、各抽出項目、ここではレース名（ｒａｃｅｎａｍｅ）、格（ｇｒａｄｅ）、競馬場（ｃｅｒｃｌｅ）、月日（ｍｍｄｄ）、距離（ｄｉｓｔａｎｃｅ）、天候・馬場（ｃｏｎｄｉｔｉｏｎ）、タイム（ｔｉｍｅ）、勝ち馬（ｗｉｎｈｏｕｒｓｅ）、性齢（ｓｅｘ＿ａｇｅ）、騎手（ｊｏｃｋｅｙ）、調教師（ｔｅｋｉ）、リンク先（ｕｒｌ）の各項目と、これら各項目を抽出するための抽出テキスト規定部分とが記述されている。図４６は、図４３，図４４のＨＴＭＬ文書から図４５のテンプレートファイル１７１を用いて項目抽出（検索）を行った結果の一表示例を示す。この表示例はアプリケーションプログラム３側で３つの項目（「騎手」「勝ち馬」「レース名」）を検索項目として指定または選択した場合を示す。

【０２１１】次に、図４０、図４７～図５０を参照して、第３の実施形態の変形例を説明する。第３の実施形態では、図４０に示すように同一ＨＴＭＬ文書内の第１および第２の表は、同一の構成要素からなる２つの部分構造に対応してテンプレートが定義されている。尚、ここで部分構造とは、例えば表、リスト、箇条書きなどで表現される意味のある１つのデータ群をいう。一方この変形例は、第１に同一ＨＴＭＬ文書内の任意の項目が他の項目と異なる属性情報を含む場合にも対応できるテンプレートを用いて上記の項目抽出を行う点、第２に同一ＨＴＭＬ文書内の異なる項目からなる複数の部分構造に対応できるテンプレートを用いて上記の項目抽出を行う点、第３にＨＴＭＬ文書がリンクを含む場合にも対応できるテンプレートを用いて上記の項目抽出を行う点において、第３の実施形態の変形である。

【０２１２】図４７、図４８は、店名情報を示すＨＴＭＬ文書のWebブラウザによる表示例を示す。図４７と図４８とは、それぞれ３つの表からなり、同様の文書構造を有するＨＴＭＬ文書である。図４９は、図４７の表示に対応するＨＴＭＬ記述を、図５０は、図４８の表示に対応するＨＴＭＬ記述を示す。図４０は、図４７および図４８（図４９および図５０）から項目を抽出するためのテンプレートを示す。図４０のテンプレートは、表または箇条書きなどのＨＴＭＬ文書中の部分構造の終端（TableEndDelimiter）、抽出項目名（Word）、抽出項目のデータ型（Word）、抽出テキスト規定部（HtmlTemplate）に関する記述からなる。例えば、ＨＴＭＬ文書中の＜／ＴＡＢＬＥ＞の出現を部分構造の終端とすることを、TableEndDelimiter=＜／ＴＡＢＬＥ＞と記述する。

【０２１３】図４９が示す<A HREF="./html_2.html">は、図５０のＨＴＭＬ文書へのリンクを示す。テンプレート解析部１３４１は、このリンク情報を解析する。テンプレート処理部１３４３は、このリンク情報に従い図４０のテンプレートの記述（NextURL）に基づいて、図４７のＨＴＭＬ文書のみでなく図４８のＨＴＭＬ文書からテンプレートを参照して項目抽出を行う。

【０２１４】図４７の表示に対応する図４９のＨＴＭＬ記述中第１の表と第２の表とは、同一構成要素の文書構造、同一表示形式の情報を備えた２つの部分構造である。テンプレート処理部１３４３は、図４０のテンプレートの第１および第２の部分構造（図５３では表）に関する記述に基づき、同一ＨＴＭＬ文書内の同一文書構造の複数の部分構造の項目情報を抽出する。図４８の表示に対応する図５０のＨＴＭＬ記述は図４９のＨＴＭＬ記述と同様の文書構造を有し、図４０のテンプレートにより図４９のＨＴＭＬソース記述と同様に項目情報が抽出される。

【０２１５】尚、図４９のＨＴＭＬソース記述中第１の表と第２の表とは、異なる属性（図４９では表示属性）を含む２つの部分構造である。図４９のＨＴＭＬ記述中構成要素「ジャンル」に対応するデータは、＜Ｉ＞と＜／Ｉ＞で囲まれた構造のものと、そうでない構造のものがある。この「Ｉ」タグは、対応するデータをイタリック書体で表示することを示す。同様に「Ｂ」タグは、対

応するデータを太字で表示することを示す。これらの異なる属性に関する情報は、図４０のテンプレート上では、同一行について２つの記述として定義されている。取得されたＨＴＭＬ文書がいずれかの行の記述に合致すれば、対応する項目が抽出される。図４０では、上記属性に対応する記述として、省略を示すタグ「．．」が用いられているので、任意の属性を有するデータを抽出することができる。

【０２１６】一方、図４７の表示に対応する図４９のＨＴＭＬソース記述中第１および第２の表に対し第３の表は、異なる抽出項目に対応する構成要素「評価」に対応するデータを含む部分構造である。テンプレート処理部１３４３は、図４０の第３の表に対応する記述に基づいて、同一ＨＴＭＬ文書内の異なる構成要素の文書構造の複数部分構造を抽出する。

【０２１７】以上説明したように、第３の実施形態によれば、複数の任意のＨＴＭＬ文書に対し、当該ＨＴＭＬ文書が内包する情報に関する各種の情報を管理し、当該情報を用いてユーザに対し適切な情報を項目別に抽出し、表形式などの統一フォーマットにて提供することが可能となる。また、ユーザが要求する抽出対象のみを抽出テキスト規定部分に規定することにより、システム構築／維持管理が容易となる。即ち、各ＨＴＭＬ文書が有する多種多様なインタフェースの相違に拘わらず、オープンなネットワーク上に散在する複数のＨＴＭＬ文書から、情報を項目別に抽出することが可能となり、抽出した情報をユーザが所望する形式により提供することが可能となる。

【０２１８】このように、第３の実施形態は、ＨＴＭＬの構文規則に依存しないテンプレートを用いて、ＨＴＭＬ文書から所望する項目を抽出する。即ち、タグまたはこれに準ずるデリミタ付きテキストであれば項目の抽出が可能である。また、抽出のための情報を定義するテンプレートファイルを作成するだけで、この項目の抽出を行う。テンプレートファイルは、対象となるＨＴＭＬ文書に基づき容易に作成可能であり、かつ視覚的にわかりやすいため、容易かつ柔軟にＨＴＭＬ文書に内在する情報の項目別の抽出を実現することができる。

【０２１９】尚、本発明は、上述した実施の形態に限定されるものではなく、その要旨を逸脱しない範囲において、種々変更することが可能である。

【０２２０】

【発明の効果】以上説明したように、本発明によれば、オープンなネットワークに散在する複数のＨＴＭＬ文書に対して該複数のＨＴＭＬ文書に内在する情報の文書構造、構成要素、表現形式等が互いに異なってもこれら複数の文書を跨った情報検索を実現し、ＨＴＭＬ記述上の差異をユーザ毎の統一形式に変換して一括して検索結果を返却することができる。

【０２２１】さらに、各ＨＴＭＬ文書が有する多種多様なインタフェースの相違に拘わらず、オープンなネットワーク上に散在する複数のＨＴＭＬ文書から、情報を項目別に抽出することが可能となり、抽出した情報をユーザが所望する形式により提供することが可能となる。

【０２２２】また、オープンなネットワークに散在する複数のサーチエンジンに対して検索を要求する場合、各サーチエンジン固有の入力フォームのオブジェクトを個別に管理することにより、異種の各サーチエンジンへのインターフェースの相違を解消して、複数の任意の入力項目に対応した柔軟な検索が可能となる。

【０２２３】従って従来に比較して、人手による多くの時間や労力が不要となり、検索効率が画期的に向上する。

【図面の簡単な説明】

【図１】本発明に係るＨＴＭＬ文書情報統合検索のユーザーの処理手順を説明する図である。

【図２】本発明に係るＨＴＭＬ文書情報統合検索装置の原理を説明する図である。

【図３】本発明の第１の実施形態に係るＨＴＭＬ文書情報統合検索装置の構成を示すブロック図である。

【図４】第１の実施形態に係るＨＴＭＬ文書メタデータ記憶部が有するテーブルの構成を説明する図である。

【図５】第１の実施形態における準備フェーズの処理手順を示すフローチャートである。

【図６】第１の実施形態における検索フェーズの処理手順を示すフローチャートである。

【図７】あるＨＴＭＬ文書における表示およびＨＴＭＬ記述の一例を示す図である。

【図８】他のＨＴＭＬ文書における表示およびＨＴＭＬ記述の一例を示す図である。

【図９】ＨＴＭＬ文書テーブルの内容を示す図である。

【図１０】図７（Ｂ）および図８（Ｂ）に対応するＨＴＭＬ文書－表マッピングテーブルの内容を示す図である。

【図１１】図７および図８に対応するＨＴＭＬ文書項目テーブルの内容を示す図である。

【図１２】ドメインテーブルの内容を示す図である。

【図１３】ユーザードメインテーブルの内容を示す図である。

【図１４】ドメイン変換関数テーブルの内容を示す図である。

【図１５】本発明の第２の実施形態に係るインターネット情報統合検索装置の構成を示すブロック図である。

【図１６】第２の実施形態に係るＨＴＭＬ文書メタデータ記憶部が有するテーブルの構成を説明する図である。

【図１７】第２の実施形態で使用される各サーチエンジンの入力フォームの例を説明する図である。

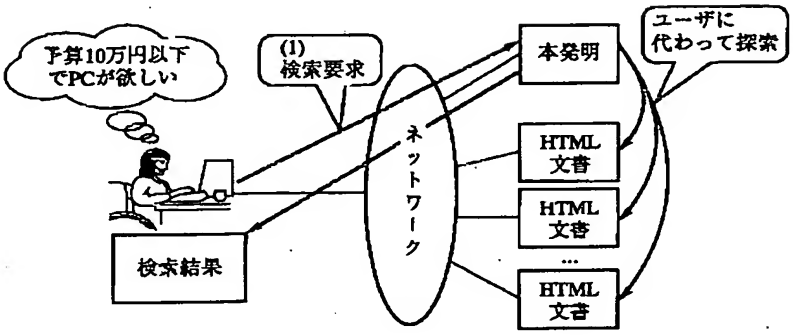
【図１８】図１７（Ｂ）の入力フォームのＨＴＭＬ記述を示す図である。

【図１９】第２の実施形態における準備フェーズの処理

手順を示すフローチャートである。
【図２０】第２の実施形態におけるＨＴＭＬ文書項目テーブルの内容の一例を説明する図である。
【図２１】第２の実施形態におけるＨＴＭＬ文書テーブルの内容の一例を説明する図である。
【図２２】第２の実施形態におけるＨＴＭＬ文書－表マッピングテーブルの内容の一例を説明する図である。
【図２３】第２の実施形態におけるドメインテーブルの内容の一例を示す図である。
【図２４】第２の実施形態におけるドメイン変換関数テーブルの内容の一例を示す図である。
【図２５】第２の実施形態におけるユーザードメインテーブルの内容の一例を示す図である。
【図２６】第２の実施形態の入力必須項目テーブルの内容の一例を示す図である。
【図２７】検索要求処理における図１５の第２の実施形態に係るインターネット情報統合検索装置と各サーチエンジンとの関係を説明する図である。
【図２８】第２の実施形態の検索パターンマトリックステーブルの内容を示す図である。
【図２９】第２の実施形態における検索フェーズの処理手順を示すフローチャートである。
【図３０】図２９のステップＳ４１０で探索されたデータ項目の所在を示す図である。
【図３１】図２９のステップＳ４４０で得られたページＡに対する検索要求の処理パターンを示す図である。
【図３２】図２９のステップＳ４４０で得られたページＢに対する検索要求の処理パターンを示す図である。
【図３３】図２９のステップＳ４４０で得られたページＣに対する検索要求の処理パターンを示す図である。
【図３４】図２９のステップＳ４５０で得られたユーザー入力ドメインとローカルドメインとの対応情報を示す図である。
【図３５】ページＢに対する検索要求の処理結果の表示例およびＨＴＭＬ記述を示す図である。
【図３６】図２９のステップＳ５００で得られたユーザー出力ドメインとローカルドメインとの対応情報を示す図である。
【図３７】本発明の第３の実施形態に係るＨＴＭＬ文書情報抽出装置の構成を示すブロック図である。
【図３８】第３の実施形態における準備フェーズの処理手順を示すフローチャートである。
【図３９】ＰＲＯＸＹ設定ファイルの内容の一例を示す図である。
【図４０】第３の実施形態におけるテンプレートファイルの内容の一例を示す図である。
【図４１】ＵＲＬ－テンプレート対応表の内容の一例を示す図である。
【図４２】第３の実施形態における実行フェーズの処理手順を示すフローチャートである。
【図４３】ＨＴＭＬ文書のWebブラウザによる表示の一例を示す図である。
【図４４】図４３の表示に対応するＨＴＭＬ記述の一部を示す図である。
【図４５】図４３、図４４に対応するテンプレートファイルの内容を示す図である。
【図４６】第３の実施形態が図４３のＨＴＭＬ文書から抽出した検索結果の表示の一例を示す図である。
【図４７】第３の実施形態の変形例におけるＨＴＭＬ文書のWebブラウザによる表示の一例を示す図である。
【図４８】図４７のＨＴＭＬ文書からリンクされる図４７の文書と同一の文書構造を有するＨＴＭＬ文書のWebブラウザによる表示の一例を示す図である。
【図４９】図４７の表示に対応するＨＴＭＬ記述を示す図である。
【図５０】図４８の表示に対応するＨＴＭＬ記述を示す図である。
【図５１】従来のＨＴＭＬ文書情報検索のユーザーの処理手順を説明する図である。
【図５２】従来のＨＴＭＬ文書情報検索の原理を説明する図である。
【符号の説明】
１　ＨＴＭＬ文書情報統合検索装置
２　ＰＲＯＸＹサーバ
３　アプリケーションプログラム
１０　インターネット情報統合検索装置
１１　ユーザーインターフェース部
１２　構文解析部
１３　問い合わせ処理部
１４　ＨＴＭＬ文書アクセス部
１５、１５０　ＨＴＭＬ文書メタデータ記憶部
１６　ＨＴＭＬ文書メタデータ管理部
４、２１、２０２　ＨＴＭＬ文書
２２　Ｗｅｂサーバー
２３　サーチエンジン
２４　データベース
１００　ＨＴＭＬ文書情報抽出装置
１３１　問い合わせ項目探索部
１３２　問い合わせ項目変換部
１３３　変換関数ライブラリ
１３４　ＨＴＭＬ文書処理部
１３５　検索結果変換部
１３６　入力必須項目探索部
１３７　検索パターン判定部
１３８　検索結果処理部
１３９　マトリックステーブル
１５１　ＨＴＭＬ文書テーブル
１５２　ＨＴＭＬ文書－表マッピングテーブル
１５３　ＨＴＭＬ文書項目テーブル
１５４　ドメインテーブル

155　ユーザードメインテーブル
156　ドメイン変換関数テーブル
157　入力必須項目テーブル
190、290　通信網
201　HTML文書要求
203　検索要求
204　検索結果
301　問い合わせ文
302　検索結果
1341　テンプレート解析部
1342　URL／テンプレート対応表
1343　テンプレート処理部
1345　テンプレートファイル

【図1】



【図9】

HTML文書テーブル

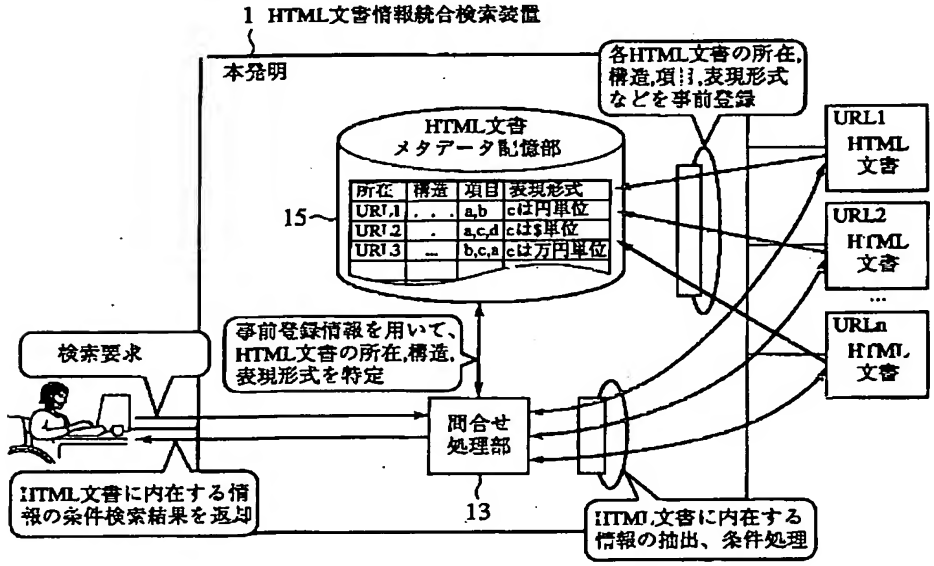
ページ名	URL
Shop_A	http://www.shop_a.co.jp/products.html
Shop_B	http://www.shop_b.co.jp/shouhin.html

【図12】

ドメインテーブル

ドメイングループ	ユーザドメイン
価格	「¥」記号つき表現形式
価格	数値と","からなる表現形式
価格	「円」記号つき表現形式

【図2】



【図10】

HTML-表マッピングテーブル

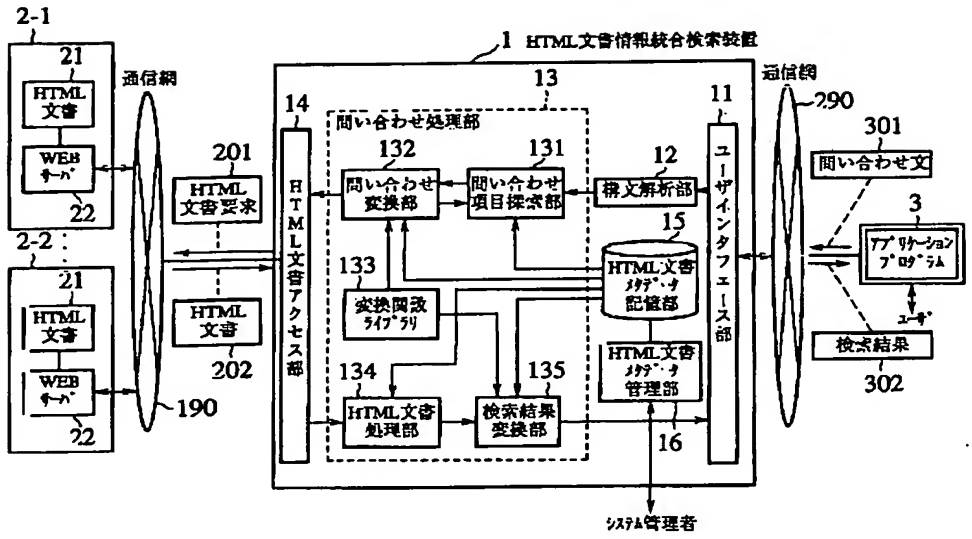
ページ名	レコード開始	列1	列2	列3	列4
Shop_A	"<TR><TD>"で始まる行	"ショップ"A"固定値	レコード開始行中の1つ目の"<TD>"と1つ目の"/"の間	レコード開始行中の1つ目の"/"と1つ目の"</TD>"の間	レコード開始行中の2つ月の"<TD>"と2つ目の"</TD>"の間
Shop_B	"<LI>"で始まる行	"ショップ"B"固定値	レコード開始行中の1つ目の"<LI>"と1つ目の"/"の間	レコード開始行中の1つ目の"/"と2つ目の"/"の間	レコード開始行中の2つ月の"/"と1つ目の"円"の間

【図13】

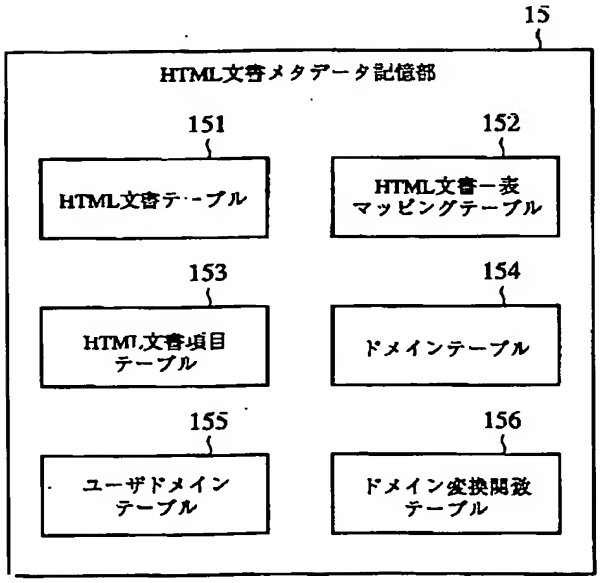
ユーザドメインテーブル

ユーザ名	ドメイングループ	ユーザ入力ドメイン	ユーザ出力ドメイン
ユーザA	価格	「円」記号つき表現形式	「円」記号つき表現形式

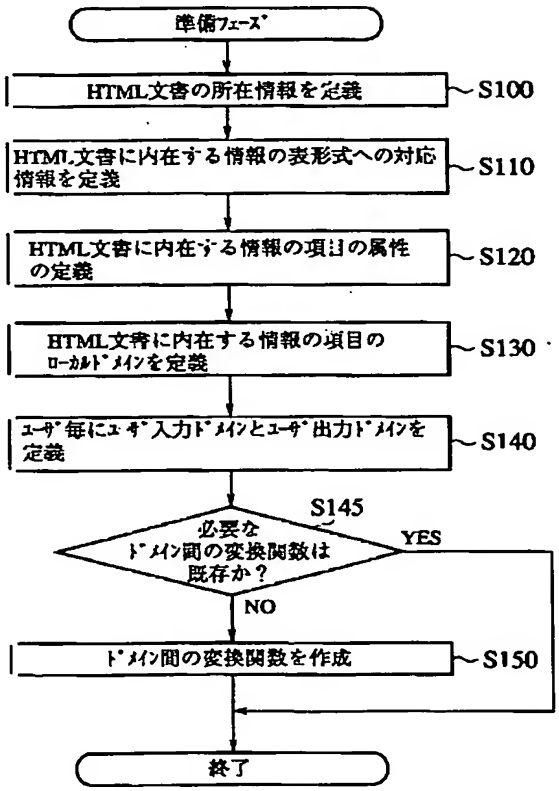
【図3】



【図4】



【図5】

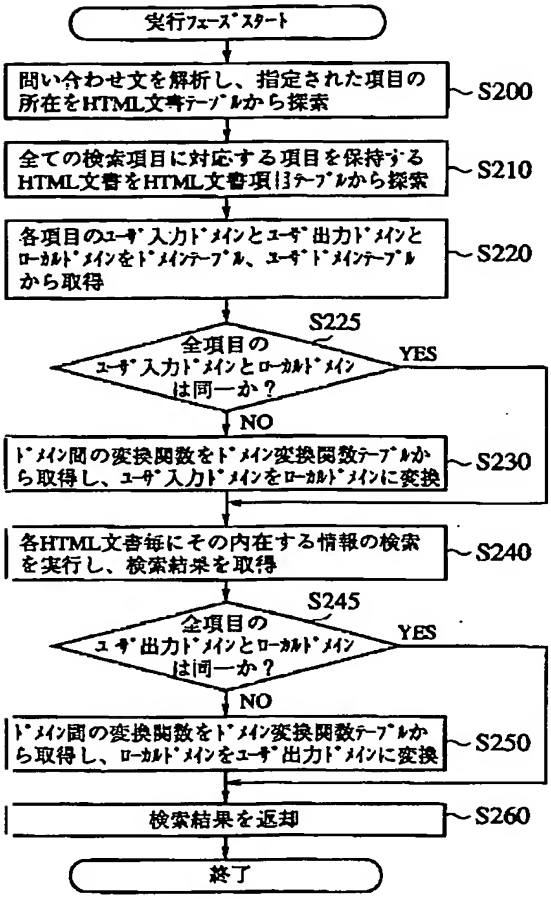


【図11】

HTML文書項目テーブル

ページ名	対応列	列名	データ型	ローカルドメイン
Shop_A	列1	ショップ名	文字列	なし
Shop_A	列2	メーカ名	文字列	なし
Shop_A	列3	商品名	文字列	なし
Shop_A	列4	価格	数値	「¥」記号つき表現形式
Shop_B	列1	ショップ名	文字列	なし
Shop_B	列2	メーカ名	文字列	なし
Shop_B	列3	商品名	文字列	なし
Shop_B	列4	価格	数値	数値と","からなる表現形式

【図6】



【図7】

HTMLの内在情報の記述にTABLEタグを用いた例

WEBブラウザでの表示例
タイトル:ショップAの取扱商品情報
URL:http://www.shop_a.co.jp/products.html

(A)

商品情報

商品名	価格
Maker A/PC1	¥170,000
Maker A/PC2	¥238,000
Maker B/PC101	¥198,000

(B) HTML文書

```
<BODY>
  <H1>商品情報 </H1>

  <TABLE BORDER>
    <TR><TH>商品名</TH><TH>価格</TH></TR>
    <TR><TD>Maker A/PC1</TD><TD>¥170,000</TD></TR>
    <TR><TD>Maker A/PC2</TD><TD>¥238,000</TD></TR>
    <TR><TD>Maker B/PC101</TD><TD>¥198,000</TD></TR>
  </TABLE>
</BODY>
```

【図14】

ドメイン変換関数テーブル

変換関数名	変換元ドメイン	変換先ドメイン	ライブラリ名
Num2Yen()	数値と","からなる表現形式	「円」記号つき表現形式	price.dll
Yen2Num()	「円」記号つき表現形式	数値と","からなる表現形式	price.dll
Num2¥()	数値と","からなる表現形式	「¥」記号つき表現形式	price.dll
¥2Num()	「¥」記号つき表現形式	数値と","からなる表現形式	price.dll
Yen2¥()	「円」記号つき表現形式	「¥」記号つき表現形式	price.dll
¥2Yen()	「¥」記号つき表現形式	「円」記号つき表現形式	price.dll

【図8】

HTMLの内在情報の記述にOLタグを用いた例

WEBブラウザでの表示例
タイトル:ショップBの取扱商品情報
URL:http://www.shop_b.co.jp/shouhin.html

(A)

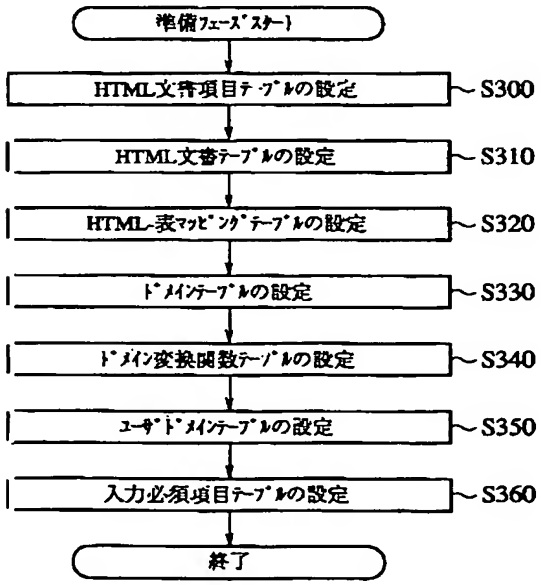
商品情報

メーカ名/商品名/価格
1. Maker A/PC1/168,000円
2. Maker B/PC101/208,000円
3. Maker B/PC102/248,000円

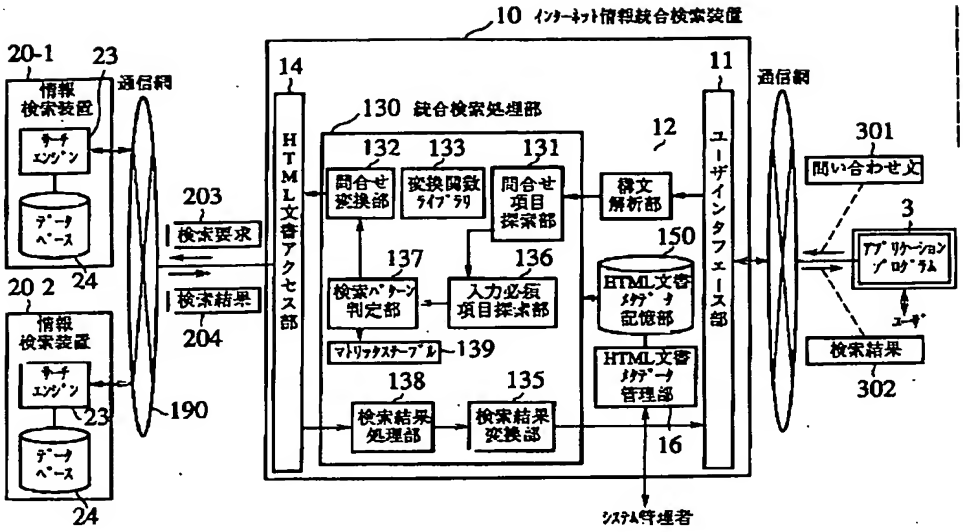
(B) HTML文書

```
<BODY>
  <H1>商品情報 </H1>

  <H3>メーカ名/商品名/価格 </H3>
  <OL>
    <LI>Maker A/PC1/168,000円
    <LI>Maker B/PC101/208,000円
    <LI>Maker B/PC102/248,000円
  </OL>
</BODY>
```

【図19】



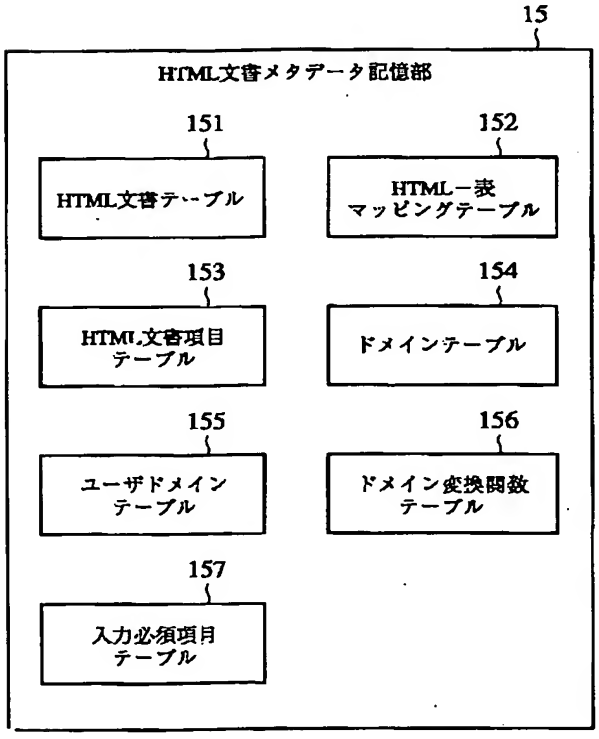
【図15】



【図30】

データ項目名−ページ名テーブル	
データ項目名	ページ名
店名	Page_A, Page_B, Page_C
電話番号	Page_A, Page_B, Page_C
エリア	Page_A, Page_B, Page_C
ジャンル	Page_A, Page_B, Page_C

【図16】



【図18】

Page_BのHTML.

```
<!DOCTYPE HTML PUBLIC"-//W3C//DTD W3 HTML//EN">
<HTML>
<HEAD>
<TITLE>page_B</TITLE>
</HEAD>
<BODY>
<FORM action=http://www.page_b.co.jp/cgi-bin/search.cgi method=GET>
検索したい店名<BR><BR>
<TABLE>
<TR>
<TH>
店名
</TH>
<TH>
エリア
</TH>
</TR>
<TR>
<TD>
<INPUT name=key size=30>
</TD>
<TD>
<SELECT name=area>
  <OPTION value=00 SELECTED>
  <OPTION value=01>横須賀市
  <OPTION value=02>藤沢市
  <OPTION value=03>平塚市
  <OPTION value=04>厚木市
  <OPTION value=05>逗子市
  <OPTION value=06>相模原市
  <OPTION value=07>横浜市
  <OPTION value=08>茅ヶ崎市
</SELECT>
</TD>
</TR>
</TABLE><BR>
<INPUT type=submit value=検索>
<INPUT type=reset value=クリア>
</FORM>
</BODY>
</HTML>
```

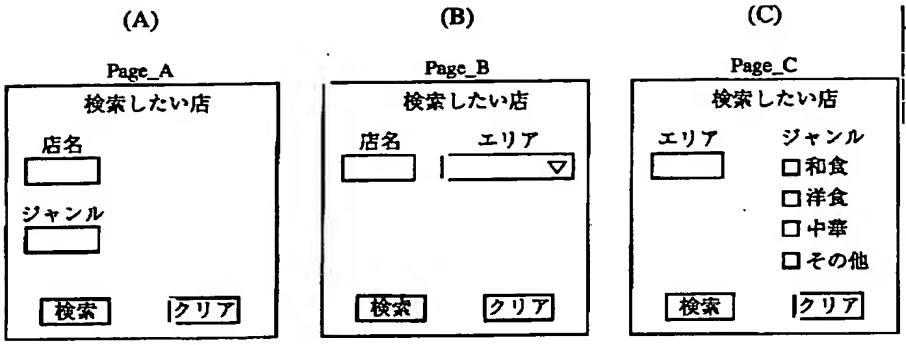
【図21】

HTML文書テーブル	
ページ名	情報検索装置URL
Page_A	http://www.Page_a.co.jp/search-shop.cgi
Page_B	http://www.Page_b.co.jp/search-shop.cgi
Page_C	http://www.Page_c.co.jp/search-shop.cgi

【図34】

ユーザ入力ドメインとローカルドメインテーブル			
ページ名	項目	ユーザ入力ドメイン	ローカルドメイン
Page_B	エリア	SHITSUKI	Page_B_City
Page_C	ジャンル	RYOURITSUKI	Page_C_Dishes

【図17】



【図26】

入力必須項目ﾃｰﾌﾞﾙ	
ﾍﾟｰｼﾞ名	入力必須項目
Page_A	ｼﾞｬﾝﾙ
Page_B	ｴﾘｱ
Page_C	ｴﾘｱ、ｼﾞｬﾝﾙ

【図20】

HTML文書項目ﾃｰﾌﾞﾙ							
ﾍﾟｰｼﾞ名	対応列	ﾃﾞｰﾀ項目名	項目指定可能	条件指定可能	ﾃﾞｰﾀ型	Nameタグ	ﾛｰｶﾙﾄﾞﾒｲﾝ
Page_A	列1	店名	○	○	文字列	shop	なし
Page_A	列2	ｴﾘｱ	○	-	文字列	—	なし
Page_A	列3	ｼﾞｬﾝﾙ	○	○	文字列	category	なし
Page_A	列4	電話番号	○	-	数値	—	Tel_Paren
Page_A	列5	URL	○	-	文字列	—	なし
Page_B	列1	店名	○	○	文字列	key	なし
Page_B	列2	ｴﾘｱ	○	○	文字列	area	Page_B_City
Page_B	列3	ｼﾞｬﾝﾙ	○	-	文字列	—	なし
Page_B	列4	電話番号	○	-	数値	—	Tel_Bar
Page_C	列1	店名	○	-	文字列	—	なし
Page_C	列2	ｴﾘｱ	○	○	文字列	area	なし
Page_C	列3	ｼﾞｬﾝﾙ	○	○	文字列	category	Page_C_Dishes
Page_C	列4	URL	○	-	文字列	—	なし
Page_C	列5	電話番号	○	-	数値	—	Tel_Bar

【図23】

ﾄﾞﾒｲﾝﾃｰﾌﾞﾙ		
ﾄﾞﾒｲﾝｸﾞﾙｰﾌﾟ	ﾄﾞﾒｲﾝ	説明
City	SHITSUKI	"市"つき表現形式
City	Page_B_City	Valueの値による表現形式
Dishes	RYOURITSUKI	"料理"つき表現形式
Dishes	Page_C_Dishes	Valueの値による表現形式
Tel	Tel_Bar	＊＊＊−＊＊＊−＊＊＊＊からなる形式
Tel	Tel_Paren	(＊＊＊) ＊＊＊ -＊＊＊＊からなる形式

【図39】

PROXYサーバ名	PROXYポート番号
Aiic.ntt.co.jp	80

【図41】

URL/ファイル名	テンプレート名
http:/www.aaa.co.jp/	aaa
http:/www.bbb.co.jp/	bbb
http:/www.ccc.co.jp/	ccc
http:/www.ddd.co.jp/	ddd
C:¥HTML¥eee.htm	eee

【図22】

HTML文書−表ﾏｯﾋﾟﾝｸﾞﾃｰﾌﾞﾙ						
ﾍﾟｰｼﾞ名	ﾚｺｰﾄﾞ開始	列1	列2	列3	列4	列5
Page_A	"<TR><TD>"で始まる行	ﾚｺｰﾄﾞ開始行中の1つ目の"<TD>"と1つ目の"</TD>"の間	ﾚｺｰﾄﾞ開始行中の2つ目の"<TD>"と2つ目の"</TD>"の間	ﾚｺｰﾄﾞ開始行中の3つ目の"<TD>"と3つ目の"</TD>"の間	ﾚｺｰﾄﾞ開始行中の4つ目の"<TD>"と4つ目の"</TD>"の間	ﾚｺｰﾄﾞ開始行中の5つ目の"<TD>"と5つ目の"</TD>"の間
Page_B	"<TR><TD>"で始まる行	ﾚｺｰﾄﾞ開始行中の1つ目の"<TD>"と1つ目の"</TD>"の間	ﾚｺｰﾄﾞ開始行中の2つ目の"<TD>"と2つ目の"</TD>"の間	ﾚｺｰﾄﾞ開始行中の3つ目の"<TD>"と3つ目の"</TD>"の間	ﾚｺｰﾄﾞ開始行中の4つ目の"<TD>"と4つ目の"</TD>"の間	Nil
Page_C	"<TR><TD>"で始まる行	ﾚｺｰﾄﾞ開始行中の1つ目の"<TD>"と1つ目の"</TD>"の間	ﾚｺｰﾄﾞ開始行中の2つ目の"<TD>"と2つ目の"</TD>"の間	ﾚｺｰﾄﾞ開始行中の3つ目の"<TD>"と3つ目の"</TD>"の間	ﾚｺｰﾄﾞ開始行中の4つ目の"<TD>"と4つ目の"</TD>"の間	ﾚｺｰﾄﾞ開始行中の5つ目の"<TD>"と5つ目の"</TD>""の間

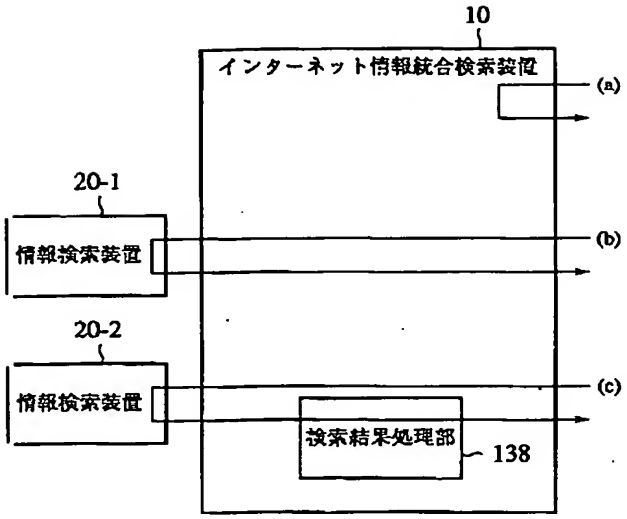
【図24】

ﾄﾞﾒｲﾝ変換関数ﾃｰﾌﾞﾙ				
変換関数名	ﾄﾞﾒｲﾝｸﾞﾙｰﾌﾟ	変換元ﾄﾞﾒｲﾝ	変換先ﾄﾞﾒｲﾝ	ﾗｲﾌﾞﾗﾘ名
Shi2ValueB ()	Area	SHITSUKI	Page_B_City	select.dll
Ryouri2ValueC ()	Category	RYOURITSUKI	Page_C_Dishes	select.dll
Bar2Paren ()	Tel	Tel_Bar	Tel_Paren	select.dll

【図25】

ﾕｰｻﾞﾄﾞﾒｲﾝﾃｰﾌﾞﾙ			
ﾕｰｻﾞ名	ﾄﾞﾒｲﾝｸﾞﾙｰﾌﾟ	ﾕｰｻﾞ入力ﾄﾞﾒｲﾝ	ﾕｰｻﾞ出力ﾄﾞﾒｲﾝ
ﾕｰｻﾞ1	Area	SHITSUKI	SHITSUKI
ﾕｰｻﾞ1	Category	RYOURITSUKI	RYOURITSUKI
ﾕｰｻﾞ1	Tel	Tel_Paren	Tel_Paren

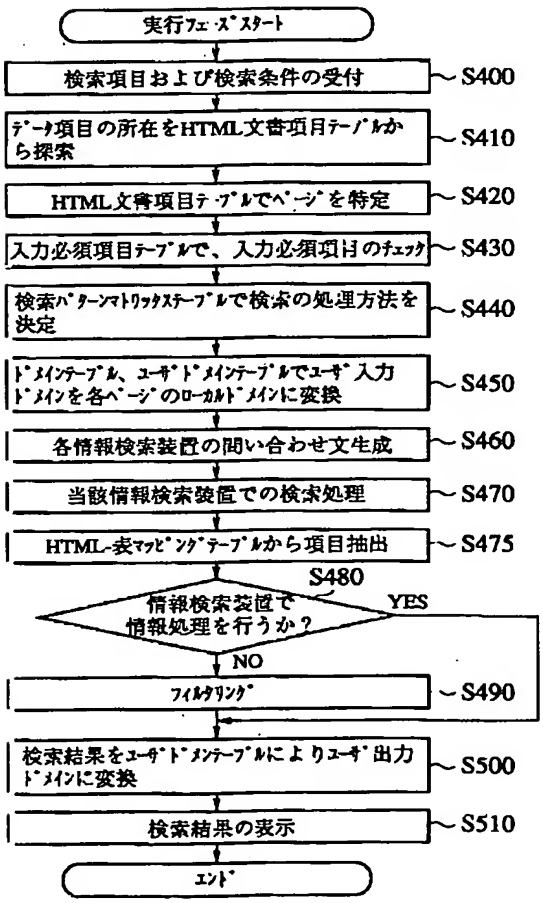
【図27】



【図28】

検索ﾊﾟﾀｰﾝﾏﾄﾘｯｸｽ							
検索要求		ｻｰﾁｴﾝｼﾞﾝ		処理			
				項目		条件	
項目	条件	項目	条件	検索条件値をそのまま返却	情報源から返却されたものを返却	ｻｰﾁｴﾝｼﾞﾝで処理	装置でﾌｨﾙﾀﾘﾝｸﾞ
○	○	○	○	×	○	○	×
○	○	○	×	○	×	×	○
○	○	×	○	○	×	○	×
○	×	○	○	×	○	×	×
○	×	○	×	×	○	×	×
○	×	×	○	−	−	−	−
×	○	○	○	×	×	○	×
×	○	○	×	×	×	×	○
×	○	×	○	×	×	○	×

【図29】



【図31】

検索要求－処理内容ﾃｰﾌﾞﾙ（ﾍﾟｰｼﾞA）								
項目名	検索要求		Page A		処理			
					項目		条件	
	項目	条件	項目	条件	検索条件値をそのまま返却	情報源から返却されたものを返却	情報検索装置で処理	装置でﾌｨﾙﾀﾘﾝｸﾞ
店名	○	×	○	○	×	○	×	×
電話番号	○	×	○	×	×	○	×	×
ｴﾘｱ	×	○	○	×	×	×	×	○
ｼﾞｬﾝﾙ	×	○	○	○	×	×	○	×

【図32】

検索要求－処理内容ﾃｰﾌﾞﾙ（ﾍﾟｰｼﾞB）								
項目名	検索要求		Page B		処理			
					項目		条件	
	項目	条件	項目	条件	検索条件値をそのまま返却	情報源から返却されたものを返却	情報検索装置で処理	装置でﾌｨﾙﾀﾘﾝｸﾞ
店名	○	×	○	○	×	○	×	×
電話番号	○	×	○	×	×	○	×	×
ｴﾘｱ	×	○	○	○	×	×	○	×
ｼﾞｬﾝﾙ	×	○	○	×	×	×	×	○

【図33】

検索要求－処理内容ﾃｰﾌﾞﾙ（ﾍﾟｰｼﾞC）								
項目名	検索要求		Page C		処理			
					項目		条件	
	項目	条件	項目	条件	検索条件値をそのまま返却	情報源から返却されたものを返却	情報検索装置で処理	装置でﾌｨﾙﾀﾘﾝｸﾞ
店名	○	×	○	×	×	○	×	×
電話番号	○	×	○	×	×	○	×	×
ｴﾘｱ	×	○	○	○	×	×	○	×
ｼﾞｬﾝﾙ	×	○	○	○	×	×	○	×

【図35】

Page_Bの検索結果のHTML

(A) WEBﾌﾞﾗｳｻﾞでの表示例

店名情報

店名	ｴﾘｱ	ｼﾞｬﾝﾙ	電話番号
B1	横浜市	和食料理	045−＊＊＊−＊＊＊＊
B2	横浜市	中華料理	045−＊＊＊−＊＊＊＊
B3	横浜市	中華料理	045−＊＊＊−＊＊＊＊

(B) HTML文書

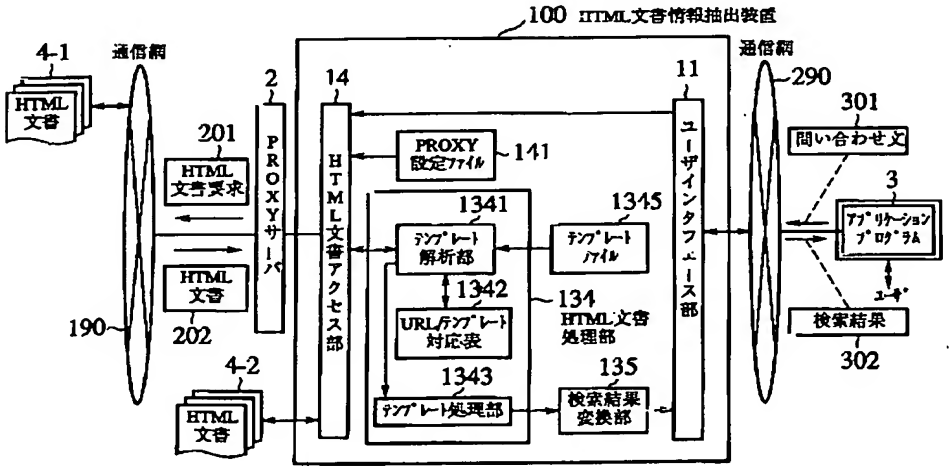
```
<BODY>
  <H1>店名情報 </H1>

  <TABLE BORDER>
    <TR><TH>店名</TH><TH>ｴﾘｱ</TH><TH>ｼﾞｬﾝﾙ</TH><TH>電話番号</TH></TR>
    <TR><TD>B1</TD><TD>横浜市</TD><TD>和食料理</TD><TD>045−＊＊＊ −＊＊＊＊</TD></TR>
    <TR><TD>B2</TD><TD>横浜市</TD><TD>中華料理</TD><TD>045−＊＊＊ −＊＊＊＊</TD></TR>
    <TR><TD>B3</TD><TD>横浜市</TD><TD>中華料理</TD><TD>045−＊＊＊−＊＊＊＊</TD></TR>
  </TABLE>
</BODY>
```

【図36】

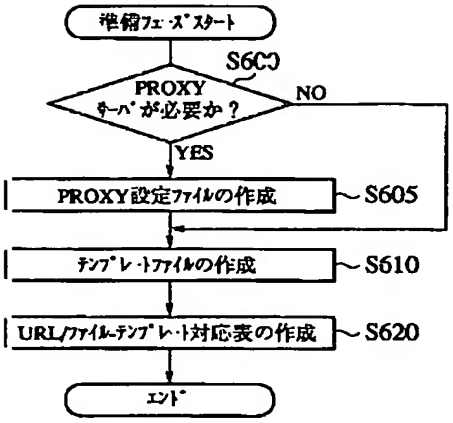
ﾕｰｻﾞ出力ﾄﾞﾒｲﾝとﾛｰｶﾙﾄﾞﾒｲﾝﾃｰﾌﾞﾙ

ﾍﾟｰｼﾞ名	項目	ﾕｰｻﾞ出力ﾄﾞﾒｲﾝ	ﾛｰｶﾙﾄﾞﾒｲﾝ
Page_A	電話番号	Tel_Paren	Tel_Paren
Page_B	電話番号	Tel_Paren	Tel_Bar
Page_C	電話番号	Tel_Paren	Tel_Bar

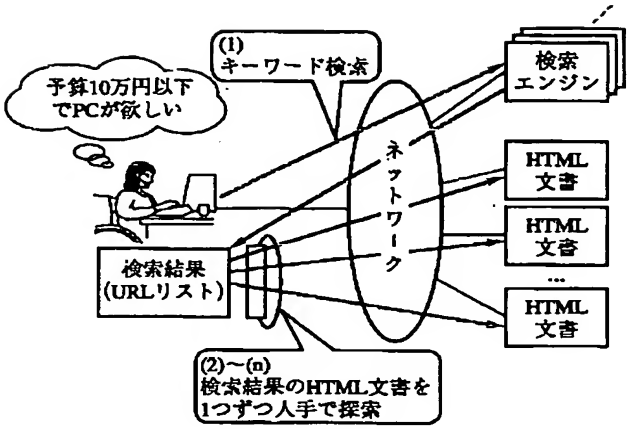
【図37】



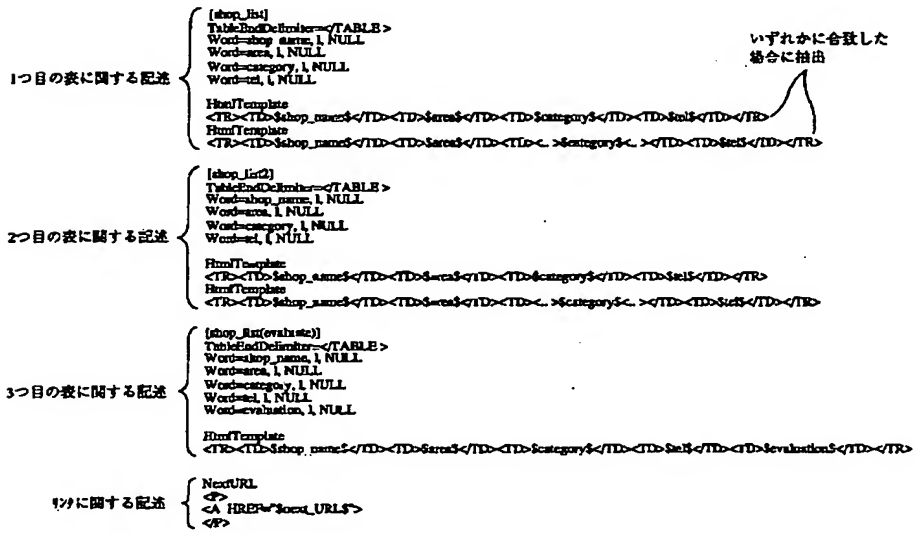
【図38】



【図51】



【図40】



【図48】

店名情報（その他）

1つ目のテーブル

店名	エリア	ジャンル	電話番号
飲食店A	鎌倉	中華料理	＊＊＊－＊＊＊－XYZA
飲食店X	横須賀	和食料理	＊＊＊－＊＊＊－YZAB
飲食店B	逗子	洋食料理	＊＊＊－＊＊＊－ABCD
飲食店Z	鎌倉	和食料理	＊＊＊－＊＊＊－WXYZ
飲食店Y	逗子	和食料理	＊＊＊－＊＊＊－QRST

店名情報（その他）

2つ目のテーブル

店名	エリア	ジャンル	電話番号
飲食店C	鎌倉	中華料理	＊＊＊－＊＊＊－XYZA
飲食店E	鎌倉	洋食料理	＊＊＊－＊＊＊－WXYZ
飲食店D	横須賀	和食料理	＊＊＊－＊＊＊－YZAB
飲食店F	逗子	洋食料理	＊＊＊－＊＊＊－ABCD
飲食店G	逗子	和食料理	＊＊＊－＊＊＊－QRST

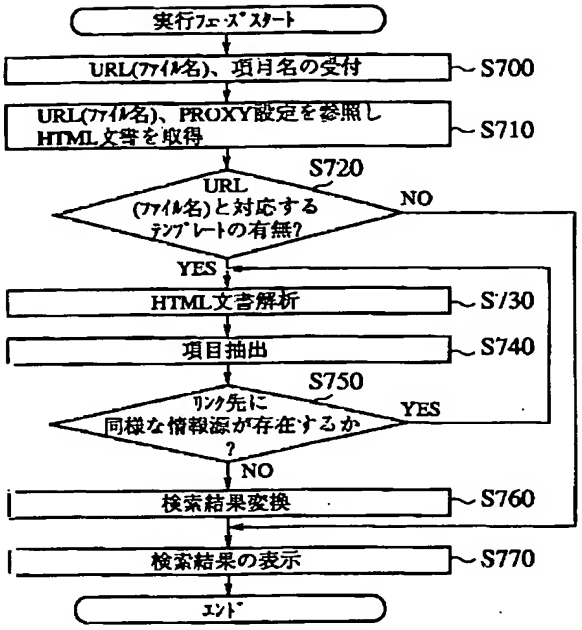
評価あり店名情報（その他）

3つ目のテーブル

店名	エリア	ジャンル	電話番号	評価
Q	鎌倉	和食料理	AABB	★★★
飲食店R	逗子	和食料理	ABAB	★★
飲食店S	川崎	中華料理	DCAB	★
飲食店T	鎌倉	洋食料理	WZAB	★★★

次ページ

【図42】



【図45】

```
Url=http://www.jba.go.jp/datafile/kachiuma//1998east.html

[JBA1998east]
Word=racename,1,NULL
Word=grade,1,NULL
Word=cercle,1,NULL
Word=mmdd,1,NULL
Word=distance,1,NULL
Word=condition,1,NULL
Word=time,1,NULL
Word=winhourse,1,NULL
Word=sex_age,1,NULL
Word=jockey,1,NULL
Word=teki,1,NULL
Word=url,5,http://www.jba.go.jp/datafile/kachiuma/

Html Template
<tr>
<_><A
HREF="$url$">"$racename$</a></td><_>$grade$</TD><_>$cercle$</TD><_>$mmdd$</TD>
<_>$distanc
e$</TD><_>$condition$</TD><_>$time$</TD><_>..</TD><_>$winhourse$</TD><_>$se
x_age$</TD><
_>$jockey$</TD><_>$teki$</TD>
</TR>
```

【図43】

平成10年産賞競走勝ち馬一覧表（関東）

レース名	格	場	月日	距離（m）	天候・馬場	タイム	勝ち馬	性齢	騎手	調教師
小山銀杯	C	小山	1. 5	2,000	曇・稍重	2:01.4	(父) メジロブロンディア	牡7	岡田 幸雄	田中 清孝
サニーネットS	C	小山	1.10	ダ1,200	曇・不良	R1:09.1	(外) スーパーカタナ	牡5	竹 豊	小西 和生
京東杯	C	小山	1.11	1,600	曇・良	1:36.8	(外) ダンツリンスキー	牡4	柴木 善臣	高橋 生二
メアリSJCC	B	小山	1.25	2,200	晴・良	2:15.3	(父) シンコウプライド	牡5	河田 洋	保見 秀一
フェアウェイS	A	東阪	2. 1	ダ1,600	曇・良	1:37.5	(父) メジロブロンディア	牡7	岡田 幸雄	田中 清孝
東阪新聞杯	C	東阪	2. 8	1,600	曇・良	1:34.2	スズホウオンナ	牡5	蛯原 正義	小島 正
東阪報告特別（春）	—	東阪	2.14	3,300	雪・良	3:39.7	ライクライクオン	牡10	御堂 信雄	義岡 八郎
同友通信杯4才S	—	東阪	2.15	ダ1,600	雪・不良	1:36.9	(外) コントレスムのキー	牡4	岡島羽 均	三ノ宮 敏
イナタモントS	C	東阪	2.21	3,200	雪・稍重	3:17.6	セーコイトップダン	牡6	河田 洋	戸無 秀幸
女王C	C	東阪	2.22	1,600	晴・良	1:35.2	(父) ダイスイライン	牝4	花沢 隆雄	須々木 康
小山牝馬S	C	小山	3. 2	1,800	晴・重	1:49.8	シンコウランバダ	牝6	立山 典弘	沼江 幸郎
四月賞	B	小山	3. 8	2,000	晴・良	2:01.8	ウルトラウィーク	牡4	竹 豊	黒井 寿昭
元月S	C	小山	3.14	ダ1,800	曇・良	1:51.9	ナイトブブクスタ	牝7	本橋 広喜	古井 仁
小山記念	B	小山	3.15	1,800	晴・良	1:48.6	サイキックスズナ	牡5	竹 豊	畠田 清
ワラーC	C	小山	3.21	1,800	曇・良	1:50.4	(外) ヒノキセニーティー	牡4	河田 洋	保見 秀一
アリンS	B	小山	3.22	1,800	曇・稍重	1:49.8	タキールハヤターン	牡4	蛯原 正義	米原 敬一
東東賞	B	小山	3.29	2,500	晴・良	2:34.4	シンオンジョウザン	牡9	江原 照男	矢田 照正
アケウルS	C	小山	4. 5	1,200	晴・良	1:08.5	(外) キンキンニーフェット	牡4	蛯原 正義	新賀 史生
アーセナル賞CT	C	小山	4.11	1,600	晴・稍重	1:34.3	(外) ホワイトホーク	牡5	岡田 幸雄	那波 栄
小山大関賞（春）	—	小山	4.18	4,100	曇・重	4:46.2	サクラゾルレインボー	牡9	田山 明	須々木 康
五月賞	A	小山	4.19	2,000	晴・良	2:01.3	ショウワンスタイ	牡4	立山 典弘	高田 一臣
グーテンT4才S	B	東阪	4.26	1,400	曇・重	1:22.2	(外) コントレスムのキー	牡4	岡島羽 均	三ノ宮 敏
ふきみ賞4才牝特	B	東阪	5. 2	2,000	晴・良	2:00.3	スピードキャンドゥ	牝4	蛯原 正義	伊東 雄二
若葉賞	C	東阪	5. 9	2,400	曇・良	2:27.6	ヤスケダイン	牡4	柴木 善臣	川内 研二
東王杯SC	B	東阪	5.16	1,400	晴・良	R1:20.1	(外) キタイシャトル	牡5	岡田 幸雄	沢藤 和雄
NTTマイルS	A	東阪	5.17	1,600	晴・稍重	1:33.7	(外) コントレスムのキー	牡4	岡島羽 均	三ノ宮 敏
古河大賞典	C	古河	5.17	2,000	曇・良	R1:58.2	ラウドハンター	牡6	吉本 晃	小久保 祥
おゆみ野S	C	東阪	5.23	2,100	晴・良	2:11.8	アイスムデラン	牡7	竹 豊	伊東 修司
雲の木賞	A	東阪	5.31	ダ2,400	晴・良	2:28.1	モルモアクセル	牝4	岡島羽 均	加東 敏二
ラブピーC	C	東阪	6. 6	1,800	曇・重	1:48.2	(外) ハナハハーモニー	牡6	立山 典弘	伊東 正博

【図44】

```
<HTML>
<HEAD>
<TITLE>平成10年度重賞競争勝ち馬一覧表（関東）</TITLE>
</HEAD>
<BODY BGCOLOR="#ffffff">
<P>
<DIV ALIGN=CENTER><FONT SIZE=5><B>平成10年度重賞競争勝ち馬一覧表（関東）
</B></FONT><BR></DIV>
<TABLE BORDER>
<TR>
<TH ALIGN=CENTER BGCOLOR="#80ffff"NOWRAP>ﾚｰｽ名</TH><TH ALIGN=CENTER BGCOLOR=
"#80ffff"
NOWRAP>格</TH><TH ALIGN=CENTER BGCOLOR="#80ffff"NOWRAP>場</TH><TH ALIGN=CENTER
BGCOLOR="#80ffff"NOWRAP>月日</TH><TH ALIGN=CENTER BGCOLOR="#80ffff"NOWRAP>距離
(m)
</TH><TH ALIGN=CENTER BGCOLOR="#80ffff"NOWRAP>天候・馬場</TH><TH ALIGN=CENTER
BGCOLOR="#80ffff"NOWRAP>ﾀｲﾑ</TH><TH COLSPAN=2 ALIGN=CENTER BGCOLOR="#80ffff"
NOWRAP>勝ち
馬</TH><TH ALIGN=CENTER BGCOLOR="#80ffff"NOWRAP>性齢</TH><TH ALIGN=CENTER BGCOL
OR="#80ffff"
NOWRAP>騎手</TH><TH ALIGN=CENTER BGCOLOR="#80ffff"NOWRAP>調教師</TH>
</TR>
<TR>
<TD ALIGN=LEFT BGCOLOR="#b5ffb5"><A HREF="1998east/19980105-0611.html">小山銀杯<
/A></TR><TD
ALIGN=CENTER BGCOLOR="#b5ffb5"NOWRAP>C</TD><TD BGCOLOR="#b5ffb5"NOWRAP>小山<
/TD>TD
BGCOLOR="#b5ffb5">1 5</TD><TD ALIGN=RIGHT BGCOLOR="#b5ffb5">2,000</TD>
<TD
BGCOLOR="#b5ffb5">稍・稍重</TD><TD ALIGN=RIGHT BGCOLOR="#b5ffb5">2:01.4</T
D><TD
BGCOLOR="#b5ffb5" NOWRAP>(父)</TD><TD BGCOLOR="#b5ffb5">ﾒｼﾞﾛﾌﾛﾝﾃｨｱ</TD
><TD
BGCOLOR="#b5ffb5">牡7</TD><TD BGCOLOR="#b5ffb5"NOWRAP>岡田　幸雄</TD><TD
BGCOLOR="#b5ffb5">田中　靖孝</TD>
</TR>
<TR>
<TD ALIGN=LEFT BGCOLOR="#b5ffb5"><A HREF="1998east/19980110-0611.html">ﾎﾟｰﾄｱ
ｲ5
</A></TD><TD ALIGN=CENTER BGCOLOR="#b5ffb5">C</TD><TD BGCOLOR="#b5ffb5">小山</
TD><TD
BGCOLOR="#b5ffb5">1.10</TD><TD ALIGN=RIGHT BGCOLOR="#b5ffb5">ﾀﾞ1,200</T
D><TD
BGCOLOR="#b5ffb5">晴・不良</TD>TD ALIGN=RIGHT BGCOLOR="#b5ffb5">R1:09.1<
/TD><TD
BGCOLOR="#b5ffb5" >(外)</TD><TD BGCOLOR="#b5ffb5">ｽｰﾊﾟｰｶﾗﾃﾏ</TD><TD
BGCOLOR="#b5ffb5">牡5</TD><TD BGCOLOR="#b5ffb5">竹　司</TD><TD BGCOLOR="#b5ffb5"
>小西　和生
</TD>
</TR>
<TR>
<TD ALIGN=LEFT BGCOLOR="#b5ffb5"><A HREF="1998east/19980111-0611.html">京東杯</A
></TD><TD
ALIGN=CENTER BGCOLOR="#b5ffb5">C<TD><TD BGCOLOR="#b5ffb5">小山</TD>TD BGCOLO
R="#b5ffb5">
```

【図４６】

候補選択番号：1

騎手	勝ち馬	レース名
岡田　幸雄	ﾒｸﾞﾙﾌﾛﾝﾃｨｱ	小山銀杯
竹　豊	ｽｰﾊﾟｰｶﾀﾔﾏ	ｻﾞｰﾈｯﾄS
柴木　善臣	ﾀﾞﾝｼﾝｽｷｰ	京東杯
河田　洋	ｼﾞﾒﾛﾌﾟﾗｲﾄ	ﾒｱﾘｶJCC
岡田　幸雄	ﾒｸﾞﾙﾌﾛﾝﾃｨｱ	ﾌﾟﾌｪﾌﾘｰS
蛯原　正義	ｽﾓｰﾙｻﾝﾃﾞｰ	東阪新聞杯
御盆　信雄	ﾃｲｹｲﾀｲｶﾞｰ	東阪障害特別（春）
岡島羽　均	ｺﾝﾄﾞﾙｴﾙﾊﾞｻﾞ	同共通信杯4才S
河田　洋	ﾋｰｺｲﾄｯﾌﾟﾃﾝ	ｲﾅｹﾞﾓﾝﾄﾞS
花沢　監徳	ﾀﾞｲｴｲｹｲﾝ	女王C
立山　典弘	ｼﾞﾒﾛﾗﾝﾊﾞﾀﾞ	小山牝馬S
竹　豊	ｳﾙﾄﾗｳｨｰｸ	四月賞
本橋　広喜	ﾁｬｲﾙﾄﾞﾌﾟﾗｽﾀ	五月S
竹　豊	ｻｲｷｯｸｽｽﾞｶ	小山記念
河田　洋	ﾋﾉｷﾐｭｰﾃｨｰ	ﾌﾗﾜｰC
蛯原　正義	ﾀﾘｰﾙﾊﾘｹｰﾝ	ﾌﾞﾘﾝS
江原　照男	ｼﾞﾝﾃﾝｼｮｳｸﾞﾝ	東東賞
蛯原　正義	ｷﾝｷﾊﾞｰﾌｪｸﾄ	ｱｸﾘﾙC
岡田　幸雄	ﾎﾜｲﾄｽﾀｰ	ﾌﾞｰﾋﾞｰ個CT
田山　剛	ｻｳｻﾞﾝﾚｲﾝﾎﾞｰ	小山大障害（春）
立山　典弘	ｼｮｳｳﾝｽｶｲ	五月賞
岡島羽　均	ｺﾝﾄﾞﾙｴﾙﾊﾞｰ	ｼﾞｰﾗﾝﾄﾞT4才S
蛯原　正義	ｽﾋﾟｰﾄﾞｷｬﾝﾄﾞｩ	ｽﾎﾟﾝｻ賞4才牝特
柴木　善臣	ﾔｽﾀｹﾞｲﾝ	若葉賞
岡田　幸雄	ｷﾀｲｼｬﾄﾙ	東王杯SC
岡島羽　均	ｺﾝﾄﾞﾙｴﾙﾊﾞｰ	NTTﾏｲﾙS
吉本　昌	ﾗｳﾄﾞﾊﾝﾀｰ	古潭大賞典
竹　豊	ｱｲｴﾑﾌﾟﾗﾝ	おゆみ野S
岡島羽　均	ﾓﾙﾓｱｸｾﾙ	桂の木賞
立山　典弘	ﾊｶﾊﾞﾊｰﾓﾆｰ	ﾗﾌﾞﾋﾞｰC

【図４７】

店名情報（横浜）

1つ目の表

店名	ｴﾘｱ	ジャンル	電話番号
和食屋1	横浜	和食料理	＊＊＊−＊＊＊＊ -XXXX
洋食屋1	横浜	洋食料理	＊＊＊−＊＊＊＊ -YYYY
和食屋2	横浜	和食料理	＊＊＊−＊＊＊＊ -ZZZZ
中華屋1	横浜	中華料理	＊＊＊−＊＊＊＊ -WWWW
和食屋3	横浜	和食料理	＊＊＊−＊＊＊＊ -QQQQ

店名情報（横浜2）

2つ目の表

店名	エリア	ジャンル	電話番号
洋食店1	横浜	洋食料理	＊＊＊−＊＊＊＊−SSSS
和食店1	横浜	和食料理	＊＊＊−＊＊＊＊−RRRR
和食店2	横浜	和食料理	＊＊＊−＊＊＊＊−TTTT
中華店1	横浜	中華料理	＊＊＊−＊＊＊＊−UUUU
洋食店2	横浜	洋食料理	＊＊＊−＊＊＊＊−VVVV

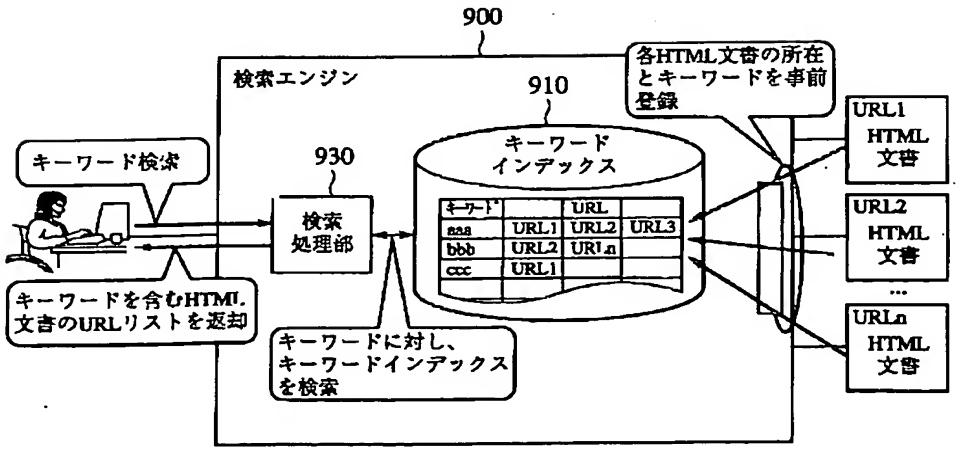
評価あり店名情報（横浜）

3つ目の表

店名	エリア	ジャンル	電話番号	評価
和食店1	横浜	和食料理	AAAA	★
洋食店1	横浜	洋食料理	BBBB	★
中華店1	横浜	中華料理	CCCC	★★★
和食店2	横浜	和食料理	DDDD	★★

リンク→次ページ

【図５２】



【図49】

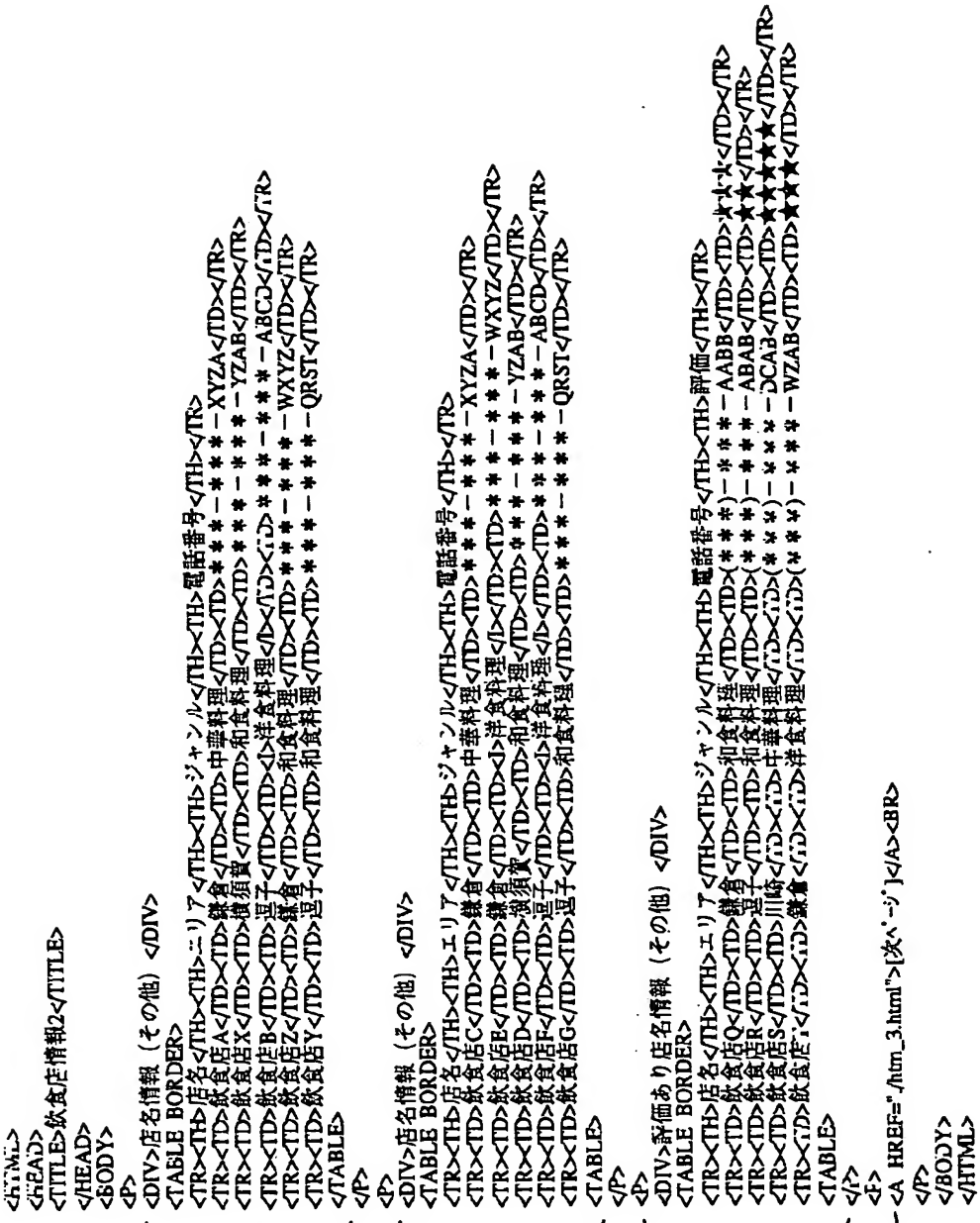
```
<HTML>
<HEAD>
<TITLE>飲食店情報1</TITLE>
</HEAD>
<BODY>
<P>
<DIV>店名情報（横浜）</DIV>
<TABLE BORDER>
<TR><TH>店名</TH><TH>エリア</TH><TH>ジャンル</TH><TH>電話番号</TH></TR>
<TR><TD>和食屋1</TD><TD>横浜</TD><TD>和食料理</TD><TD>＊＊＊－＊＊＊－XXXX</TD></TR>
<TR><TD>洋食屋1</TD><TD>横浜</TD><TD>洋食料理</TD><TD>＊＊＊－＊＊＊－YYYY</TD></TR>
<TR><TD>和食屋2</TD><TD>横浜</TD><TD>和食料理</TD><TD>＊＊＊－＊＊＊－ZZZZ</TD></TR>
<TR><TD>中華料理1</TD><TD>横浜</TD><TD>中華料理</TD><TD>＊＊＊－＊＊＊－WWWW</TD></TR>
<TR><TD>和食屋3</TD><TD>横浜</TD><TD>和食料理</TD><TD>＊＊＊－＊＊＊－QQQQ</TD></TR>
</TABLE>
</P>
<P>
<DIV>店名情報（横浜2）</DIV>
<TABLE BORDER>
<TR><TH>店名</TH><TH>エリア</TH><TH>ジャンル</TH><TH>電話番号</TH></TR>
<TR><TD>洋食店1</TD><TD>横浜</TD><TD>洋食料理</TD><TD>＊＊＊－＊＊＊－SSSS</TD></TR>
<TR><TD>和食店1</TD><TD>横浜</TD><TD>和食料理</TD><TD>＊＊＊－＊＊＊－RRRR</TD></TR>
<TR><TD>和食店2</TD><TD>横浜</TD><TD>和食料理</TD><TD>＊＊＊－＊＊＊－TTTT</TD></TR>
<TR><TD>中華店1</TD><TD>横浜</TD><TD>中華料理</TD><TD>＊＊＊－＊＊＊－UUUU</TD></TR>
<TR><TD>洋食店2</TD><TD>横浜</TD><TD>洋食料理</TD><TD>＊＊＊－＊＊＊－VVVV</TD></TR>
</TABLE>
</P>
<P>
<DIV>評価あり店名情報（横浜）</DIV>
<TABLE BORDER>
<TR><TH>店名</TH><TH>エリア</TH><TH>ジャンル</TH><TH>電話番号</TH><TH>評価</TH></TR>
<TR><TD>和食店1</TD><TD>横浜</TD><TD>和食料理</TD><TD>(＊＊＊)－＊＊＊－AAAA</TD><TD>★</TD></TR>
<TR><TD>洋食店1</TD><TD>横浜</TD><TD>洋食料理</TD><TD>(＊＊＊)－＊＊＊－BBBB</TD><TD>★★★★</TD></TR>
<TR><TD>中華店1</TD><TD>横浜</TD><TD>中華料理</TD><TD>(＊＊＊)－＊＊＊－CCCC</TD><TD>★★★</TD></TR>
<TR><TD>和食店2</TD><TD>横浜</TD><TD>和食料理</TD><TD>(＊＊＊)－＊＊＊－DDDD</TD><TD>★★</TD></TR>
</TABLE>
</P>
<P>
リンク ―― <A HREF="./htm_2.html">[次ページ]</A><BR>
</P>
</BODY>
</HTML>
```

1つ目の表

2つ目の表

3つ目の表

【図50】



【手続補正書】
【提出日】平成12年4月25日（2000．4．25）
【手続補正1】
【補正対象書類名】明細書
【補正対象項目名】特許請求の範囲
【補正方法】変更
【補正内容】
【特許請求の範囲】
【請求項1】　オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合

検索装置であって、
オープンネットワーク上での半構造化文書の所在を示す所在情報と、前記半構造化文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するために用いるデータ属性を規定するデータ属性情報と、ユーザーの表示における項目の表現形式、各半構造化文書の項目の表現形式およびこれらの間の表現形式を変換するために用いる関数を定義する表現形式変換情報とを、各半構造化文書の項目情報を記述するために参照されるメタデータとして記憶する記憶部と、
検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を前記所在情報から得る文書所在探索部と、
入力された前記問い合わせを、前記表現形式変換情報に基づいて、前記得られた所在にある半構造化文書中の前記検索項目に対応する項目の表現形式に必要に応じ前記関数を参照して変換する問い合わせ変換部と、
前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得する文書検索部と、
取得された各半構造化文書から、前記文書構造情報に基づいて、項目データを抽出し、前記検索条件を用い、前記データ属性情報に基づいて前記抽出された項目データを選択して検索結果とする文書処理部と、
前記検索結果を、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に必要に応じ前記関数を参照して変換する検索結果変換部とを具備することを特徴とする半構造化文書情報統合検索装置。

【請求項２】　上記半構造化文書情報統合検索装置は、さらに、
半構造化文書ごとに、半構造化文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、半構造化文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部と、
取得された半構造化文書に対応するテンプレートを解析するテンプレート解析部と、
前記取得された半構造化文書をスキャンして、該半構造化文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理部とを具備し、
前記テンプレートには、各項目データに対応する変数名が記述されるとともに、半構造化文書が複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項１に記載の半構造化文書情報統合検索装置。

【請求項３】　前記文書処理部は、前記検索結果を、表形式に整形することを特徴とする請求項１に記載の半構造化文書情報統合検索装置。

【請求項４】　前記文書処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他の半構造化文書へのリンク情報を含む場合には、リンク先の半構造化文書をスキャンして、前記リンク先の半構造化文書と前記テンプレートとを比較することを特徴とする請求項２に記載の半構造化文書情報統合検索装置。

【請求項５】　前記テンプレートは、半構造化文書の各部分構造に対して、前記部分構造の一部に存在する、前記文書構造情報が他の部分と異なる部分をそれぞれ抽出するための、異なるタグにそれぞれ対応する複数の抽出テキスト形式情報が記述され、
前記テンプレート処理部は、前記取得された第１の検索結果である半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項２に記載の半構造化文書情報統合検索装置。

【請求項６】　前記テンプレートは、半構造化文書が互いに異なる要素からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項２に記載の半構造化文書情報統合検索装置。

【請求項７】　オープンネットワーク上の複数のサーチエンジンを介して情報を検索する半構造化文書情報統合検索装置であって、
オープンネットワーク上でのサーチエンジンの所在を示す所在情報と、各サーチエンジンへの入力フォームにおいて入力が必要とされる入力必須項目を定義する入力必須項目情報と、ＨＴＭＬ文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、項目ごとに各サーチエンジン内において該項目が取得可能か否かおよび条件指定可能か否かを示す項目属性情報と、前記項目ごとに前記項目を条件検索するためのデータ属性を規定するデータ属性情報と、ユーザーの項目の表現形式と各ＨＴＭＬ文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶する記憶部と、
検索項目および検索条件からなるユーザーから入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を前記所在情報から得る文書所在探索部と、
前記入力必須項目情報に基づいて、各サーチエンジンにおける入力必須項目と前記入力された問い合わせで指定された項目とを比較することにより、前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索項目が指定されたサーチエンジンを、検索対象サー

チエンジンとして選択するサーチエンジン選択部と、
前記入力された検索項目および検索条件と、前記項目属性情報との組み合わせを規定するマトリックステーブルに基づき各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定部と、
前記変換された問い合わせ群のそれぞれを、前記表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換部と、
前記変換された問い合わせを前記得られた所在に送信して、ＨＴＭＬ文書を取得する文書検索部と、
各サーチエンジンから取得されたＨＴＭＬ文書からなる第１の検索結果から、前記文書構造情報に基づいて、項目データを抽出するとともに、少なくともサーチエンジンにおいて条件検索が実行されなかった項目に関し、対応する前記検索処理パターンに従い、前記検索条件および前記データ属性情報に基づいて、抽出された項目データから前記検索条件に合致する項目データを選択して、第２の検索結果とする検索結果処理部と、
前記第２の検索結果を、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換部とを具備することを特徴とする半構造化文書情報統合検索装置。
【請求項８】　上記半構造化文書情報統合検索装置は、さらに、
ＨＴＭＬ文書ごとに、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部と、
取得されたＨＴＭＬ文書に対応するテンプレートを解析するテンプレート解析部と、
前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理部とを具備し、
前記テンプレートには、各項目データに対応する変数名が記述されるとともに、ＨＴＭＬ文書が複数の同一部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項７に記載の半構造化文書情報統合検索装置。
【請求項９】　前記文書処理部は、前記検索結果を、表形式に整形することを特徴とする請求項７に記載の半構造化文書情報統合検索装置。
【請求項１０】　前記文書処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他のＨＴＭＬ文書へのリンク情報を含む場合には、リンク先のＨＴＭＬ文書をスキャンして、前記リンク先のＨＴＭＬ文書と前記テンプレートとを比較することを特徴とする請求項８に記載の半構造化文書情報統合検索装置。
【請求項１１】　前記テンプレートは、ＨＴＭＬ文書の各部分構造に対して、前記部分構造の一部に存在する、前記文書構造情報が他の部分と異なる部分をそれぞれ抽出するための、異なるタグにそれぞれ対応する複数の抽出テキスト形式情報が記述され、
前記テンプレート処理部は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項８に記載の半構造化文書情報統合検索装置。
【請求項１２】　前記テンプレートは、ＨＴＭＬ文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記文書処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項８に記載の半構造化文書情報統合検索装置。
【請求項１３】　オープンネットワーク上の任意のＨＴＭＬ文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出装置であって、
ＨＴＭＬ文書ごとに、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶部と、
取得されたＨＴＭＬ文書に対応するテンプレートを解析するテンプレート解析部と、
前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理部とを具備し、
前記テンプレートには、各項目データに対応する変数名が記述されるとともに、ＨＴＭＬ文書が複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする半構造化文書情報抽出装置。
【請求項１４】　前記テンプレート処理部は、前記抽出された項目データを、表形式に整形することを特徴とする請求項１３に記載の半構造化文書情報抽出装置。
【請求項１５】　前記テンプレート処理部は、前記テンプレート中の前記抽出テキスト形式情報が、他のＨＴＭ

L文書へのリンク情報を含む場合には、リンク先のHTML文書をスキャンして、前記リンク先のHTML文書と前記テンプレートとを比較することを特徴とする請求項１３に記載の半構造化文書情報抽出装置。

【請求項１６】　前記テンプレートは、HTML文書の各部分構造に対して、前記部分構造の一部に存在する、前記文書構造情報が他の部分と異なる部分をそれぞれ抽出するための、異なるタグにそれぞれ対応する複数の抽出テキスト形式情報が記述され、

前記テンプレート処理部は、前記取得された第１の検索結果であるHTML文書をスキャンして、該HTML文書と、該HTML文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項１３に記載の半構造化文書情報抽出装置。

【請求項１７】　前記テンプレートは、HTML文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記テンプレート処理部は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項１３に記載の半構造化文書情報抽出装置。

【請求項１８】　オープンネットワーク上の複数の半構造化文書に内在する情報を検索する半構造化文書情報統合検索方法であって、

オープンネットワーク上での半構造化文書の所在を示す所在情報と、前記半構造化文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するために用いるデータ属性を規定するデータ属性情報と、ユーザーの表示における項目の表現形式、各半構造化文書の項目の表現形式およびこれらの間の表現形式を変換するために用いる関数を定義する表現形式変換情報とを、各半構造化文書の項目情報を記述するために参照されるメタデータとして記憶するステップと、

検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を前記所在情報から得るステップと、

入力された前記問い合わせを、前記表現形式変換情報に基づいて、前記得られた所在にある半構造化文書中の前記検索項目に対応する項目の表現形式に必要に応じ前記関数を参照して変換するステップと、

前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得するステップと、

取得された各半構造化文書から、前記文書構造情報に基づいて、項目データを抽出し、前記検索条件を用い、前記データ属性情報に基づいて前記抽出された項目データを選択して検索結果とするステップと、

前記検索結果を、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に必要に応じ前記関数を参照して変換するステップとを含むことを特徴とする半構造化文書情報統合検索方法。

【請求項１９】　オープンネットワーク上の複数のサーチエンジンを介して情報を検索する半構造化文書情報統合検索方法であって、

オープンネットワーク上でのサーチエンジンの所在を示す所在情報と、各サーチエンジンへの入力フォームにおいて入力が必要とされる入力必須項目を定義する入力必須項目情報と、HTML文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、項目ごとに各サーチエンジン内において該項目が取得可能か否かおよび条件指定可能か否かを示す項目属性情報と、前記項目ごとに前記項目を条件検索するためのデータ属性を規定するデータ属性情報と、ユーザーの項目の表現形式と各HTML文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶するステップと、

検索項目および検索条件からなるユーザーから入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を前記所在情報から得るステップと、

前記入力必須項目情報に基づいて、各サーチエンジンにおける入力必須項目と前記入力された問い合わせで指定された項目とを比較することにより、前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索項目が指定されたサーチエンジンを、検索対象サーチエンジンとして選択するステップと、

前記入力された検索項目および検索条件と、前記項目属性情報との組み合わせを規定するマトリックステーブルに基づき各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換するステップと、

前記変換された問い合わせ群のそれぞれを、前記表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換するステップと、

前記変換された問い合わせを前記得られた所在に送信して、HTML文書を取得するステップと、

各サーチエンジンから取得されたHTML文書からなる第１の検索結果から、前記文書構造情報に基づいて、項目データを抽出するとともに、少なくともサーチエンジンにおいて条件検索が実行されなかった項目に関し、対応する前記検索処理パターンに従い、前記検索条件および前記データ属性情報に基づいて、抽出された項目データから前記検索条件に合致する項目データを選択して、第２の検索結果とするステップと、

前記第２の検索結果を、前記表現形式変換情報に基づい

て、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換するステップとを含むことを特徴とする半構造化文書情報統合検索方法。

【請求項２０】　オープンネットワーク上の任意のＨＴＭＬ文書に内在する情報から項目ごとに情報を抽出する半構造化文書情報抽出方法であって、

ＨＴＭＬ文書ごとに、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するステップと、

取得されたＨＴＭＬ文書に対応するテンプレートを解析するステップと、

前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するステップとを含み、

前記テンプレートには、各項目データに対応する変数名が記述されるとともに、ＨＴＭＬ文書が複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記項目データを抽出するステップは、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする半構造化文書情報抽出方法。

【請求項２１】　オープンネットワーク上の複数の半構造化文書に内在する情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、

オープンネットワーク上での半構造化文書の所在を示す所在情報と、前記半構造化文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、前記項目ごとに前記項目を条件検索するために用いるデータ属性を規定するデータ属性情報と、ユーザーの表示における項目の表現形式、各半構造化文書の項目の表現形式およびこれらの間の表現形式を変換するために用いる関数を定義する表現形式変換情報とを、各半構造化文書の項目情報を記述するために参照されるメタデータとして記憶する記憶処理と、

検索項目および検索条件からなる入力された問い合わせに基づいて、すべての検索項目に対応する項目を有する半構造化文書の所在を前記所在情報から得る文書所在探索処理と、

入力された前記問い合わせを、前記表現形式変換情報に基づいて、前記得られた所在にある半構造化文書中の前記検索項目に対応する項目の表現形式に必要に応じ前記関数を参照して変換する問い合わせ変換処理と、

前記変換された問い合わせを前記得られた所在に送信して、半構造化文書を取得する文書検索処理と、

取得された各半構造化文書から、前記文書構造情報に基づいて、項目データを抽出し、前記検索条件を用い、前記データ属性情報に基づいて前記抽出された項目データを選択して検索結果とする文書処理と、

前記検索結果を、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に必要に応じ前記関数を参照して変換する検索結果変換処理とを含むことを特徴とするコンピュータ読み取り可能な記録媒体。

【請求項２２】　上記コンピュータ読み取り可能な記録媒体は、さらに、

半構造化文書ごとに、半構造化文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、半構造化文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶処理と、

取得された半構造化文書に対応するテンプレートを解析するテンプレート解析処理と、

前記取得された半構造化文書をスキャンして、該半構造化文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理とを含み、

前記テンプレートには、各項目データに対応する変数名が記述されるとともに、半構造化文書が複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項２１に記載のコンピュータ読み取り可能な記録媒体。

【請求項２３】　前記文書処理は、前記検索結果を、表形式に整形することを特徴とする請求項２１に記載のコンピュータ読み取り可能な記録媒体。

【請求項２４】　前記文書処理は、前記テンプレート中の前記抽出テキスト形式情報が、他の半構造化文書へのリンク情報を含む場合には、リンク先の半構造化文書をスキャンして、前記リンク先の半構造化文書と前記テンプレートとを比較することを特徴とする請求項２２に記載のコンピュータ読み取り可能な記録媒体。

【請求項２５】　前記テンプレートは、半構造化文書の各部分構造に対して、前記部分構造の一部に存在する、前記文書構造情報が他の部分と異なる部分をそれぞれ抽出するための、異なるタグにそれぞれ対応する複数の抽出テキスト形式情報が記述され、

前記テンプレート処理は、前記取得された第１の検索結果である半構造化文書をスキャンして、該半構造化文書と、該半構造化文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項２２に記載のコンピュータ読み取り可能な記録媒体。

【請求項２６】　前記テンプレートは、半構造化文書が互いに異なる要素からなる複数の部分構造を有する場

合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項２２に記載のコンピュータ読み取り可能な記録媒体。
【請求項２７】　オープンネットワーク上の複数のサーチエンジンを介して情報を検索する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、
オープンネットワーク上でのサーチエンジンの所在を示す所在情報と、各サーチエンジンへの入力フォームにおいて入力が必要とされる入力必須項目を定義する入力必須項目情報と、ＨＴＭＬ文書の構造を、抽出すべき項目ごとに区切るための文書構造情報と、項目ごとに各サーチエンジン内において該項目が取得可能か否かおよび条件指定可能か否かを示す項目属性情報と、前記項目ごとに前記項目を条件検索するためのデータ属性を規定するデータ属性情報と、ユーザーの項目の表現形式と各ＨＴＭＬ文書の項目の表現形式との間の変換情報を定義する表現形式変換情報とを記憶する記憶処理と、
検索項目および検索条件からなるユーザーから入力された問い合わせに基づいて、すべての検索項目に対応する項目を有するサーチエンジンの所在を前記所在情報から得る文書所在探索処理と、
前記入力必須項目情報に基づいて、各サーチエンジンにおける入力必須項目と前記入力された問い合わせで指定された項目とを比較することにより、前記得られた所在のサーチエンジンの中から、前記入力必須項目を満たす検索項目が指定されたサーチエンジンを、検索対象サーチエンジンとして選択するサーチエンジン選択処理と、
前記入力された検索項目および検索条件と、前記項目属性情報との組み合わせを規定するマトリックステーブルに基づき各サーチエンジンごとに最適な検索処理パターンを得て、前記問い合わせを各サーチエンジンごとに前記検索処理パターンに適合する問い合わせ群に変換する検索パターン判定処理と、
前記変換された問い合わせ群のそれぞれを、前記表現形式変換情報に基づいて、前記検索対象サーチエンジンの前記検索項目に対応する項目の表現形式に変換する問い合わせ変換処理と、
前記変換された問い合わせを前記得られた所在に送信して、ＨＴＭＬ文書を取得する文書検索処理と、
各サーチエンジンから取得されたＨＴＭＬ文書からなる第１の検索結果から、前記文書構造情報に基づいて、項目データを抽出するとともに、少なくともサーチエンジン内において条件検索が実行されなかった項目に関し、対応する前記検索処理パターンに従い、前記検索条件および前記属性情報に基づいて、抽出された項目データから前記検索条件に合致する項目データを選択して、第２の検索結果とする検索結果生成処理と、
前記第２の検索結果を、前記表現形式変換情報に基づいて、前記検索結果中の項目に対応する各ユーザーごとに定義された項目の表現形式に変換する検索結果変換処理とを具備することを特徴とするコンピュータ読み取り可能な記録媒体。
【請求項２８】　上記コンピュータ読み取り可能な記録媒体は、さらに、
ＨＴＭＬ文書ごとに、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶処理と、
取得されたＨＴＭＬ文書に対応するテンプレートを解析するテンプレート解析処理と、
前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理とを含み、
前記テンプレートには、各項目データに対応する変数名が記述されるとともに、ＨＴＭＬ文書が複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、
前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項２７に記載のコンピュータ読み取り可能な記録媒体。
【請求項２９】　前記文書処理は、前記検索結果を、表形式に整形することを特徴とする請求項２７に記載のコンピュータ読み取り可能な記録媒体。
【請求項３０】　前記文書処理は、前記テンプレート中の前記抽出テキスト形式情報が、他のＨＴＭＬ文書へのリンク情報を含む場合には、リンク先のＨＴＭＬ文書をスキャンして、前記リンク先のＨＴＭＬ文書と前記テンプレートとを比較することを特徴とする請求項２８に記載のコンピュータ読み取り可能な記録媒体。
【請求項３１】　前記テンプレートは、ＨＴＭＬ文書の各部分構造に対して、前記部分構造の一部に存在する、前記部分構造情報が異なる部分をそれぞれ抽出するための異なるタグにそれぞれ対応する複数の抽出テキスト形式情報が記述され、
前記テンプレート処理は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項２８に記載のコンピュータ読み取り可能な記録媒体。
【請求項３２】　前記テンプレートは、ＨＴＭＬ文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項２８に記載のコンピュータ読み取り可能な記録媒体。

【請求項３３】　オープンネットワーク上の任意のＨＴＭＬ文書に内在する情報から項目ごとに情報を抽出する処理をコンピュータに実行させるプログラムを記録するコンピュータ読み取り可能な記録媒体であって、

ＨＴＭＬ文書ごとに、ＨＴＭＬ文書の構造を抽出すべき項目ごとに区切るための文書構造情報に基づき、少なくとも抽出すべき項目名と、ＨＴＭＬ文書から抽出すべき項目群の抽出テキスト形式情報を記述するテンプレートを記憶するテンプレート記憶処理と、

取得されたＨＴＭＬ文書に対応するテンプレートを解析するテンプレート解析処理と、

前記取得されたＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、前記テンプレートとを比較して、前記抽出テキスト形式情報に合致した項目の項目データを抽出するテンプレート処理とを含み、

前記テンプレートには、各項目データに対応する変数名が記述されるとともに、ＨＴＭＬ文書が複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とするコンピュータ読み取り可能な記録媒体。

【請求項３４】　前記テンプレート処理は、前記抽出された項目データを、表形式に整形することを特徴とする請求項３３に記載のコンピュータ読み取り可能な記録媒体。

【請求項３５】　前記テンプレート処理は、前記テンプレート中の前記抽出テキスト形式情報が、他のＨＴＭＬ文書へのリンク情報を含む場合には、リンク先のＨＴＭＬ文書をスキャンして、前記リンク先のＨＴＭＬ文書と前記テンプレートとを比較することを特徴とする請求項３３に記載のコンピュータ読み取り可能な記録媒体。

【請求項３６】　前記テンプレートは、ＨＴＭＬ文書の各部分構造に対して、前記部分構造の一部に存在する、前記文書構造情報が他の部分と異なる部分をそれぞれ抽出するための、異なるタグにそれぞれ対応する複数の抽出テキスト形式情報が記述され、

前記テンプレート処理は、前記取得された第１の検索結果であるＨＴＭＬ文書をスキャンして、該ＨＴＭＬ文書と、該ＨＴＭＬ文書に対応する前記テンプレート中の前記複数の抽出テキスト形式情報のいずれかが合致した場合に、合致した項目の項目データを抽出することを特徴とする請求項３３に記載のコンピュータ読み取り可能な記録媒体。

【請求項３７】　前記テンプレートは、ＨＴＭＬ文書が互いに異なる項目からなる複数の部分構造を有する場合、各部分構造ごとに抽出テキスト形式情報が記述され、

前記テンプレート処理は、抽出された項目データを、各部分構造ごとの検索結果とすることを特徴とする請求項３３に記載のコンピュータ読み取り可能な記録媒体。

フロントページの続き

(72)発明者　永末　壽宏
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

(72)発明者　星野　隆
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

(72)発明者　町原　宏毅
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

Ｆターム(参考)　5B075 KK02 KK03 KK07 ND34 QM05
5B082 GA02 GC04